## Conference Abstracts

Listed by author (alphabetical)

### KEYNOTE SPEAKERS

<table>
<thead>
<tr>
<th>Author</th>
<th>Paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conrad, Susan</td>
<td>From a plate of spaghetti to a cable–stayed bridge: increasing the impact of corpus linguistics in disciplinary education</td>
</tr>
<tr>
<td>Hardie, Andrew</td>
<td>Exploratory analysis of word frequencies across corpus texts: towards a critical contrast of approaches</td>
</tr>
<tr>
<td>Mair, Christian</td>
<td>Downsizing and upgrading: why we need more spoken, more multilingual and more nonstandard corpora</td>
</tr>
<tr>
<td>McIntyre, Dan</td>
<td>Just what is corpus stylistics?</td>
</tr>
<tr>
<td>Scott, Mike</td>
<td>News downloads and aboutness</td>
</tr>
</tbody>
</table>

### GENERAL SPEAKERS (an asterisk * indicates poster)

<table>
<thead>
<tr>
<th>Author</th>
<th>Paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adolphs, Svenja</td>
<td>Multimodal Corpus–based Investigations of the Gesture–Speech Relationship</td>
</tr>
<tr>
<td>Afsar, Ayaz</td>
<td>The Ideological Representation of ‘self’ and ‘other’ in Post 9/11 Discourse in British Newspapers</td>
</tr>
<tr>
<td>Akbas, Erdem</td>
<td>Exploring the interaction across the texts of non–native speakers of English: Compiling reliable learner corpora to identify hedging and boosting in argumentative writing in EFL context</td>
</tr>
<tr>
<td>Akef, Alaa Mamdouh</td>
<td>A Parallel Corpus–based Study of Chinese Arabic verb phrase alignment</td>
</tr>
<tr>
<td>Alasmary, Abdullah</td>
<td>A corpus–based investigation of the lexical bundle use by accomplished and novice mathematics writers</td>
</tr>
<tr>
<td>Albrecht, Sven</td>
<td>Corpus–Linguistic and Cognitive Approaches to Determiner Usage in Chinese Student Writing Testing the Fluctuation Hypothesis</td>
</tr>
<tr>
<td>Alfraidi, Tareq</td>
<td>Arabic Concessive Conditional Particles and Modality Meanings: A corpus–based study (*)</td>
</tr>
<tr>
<td>Algouzi, Sami</td>
<td>Functions of discourse marker ‘you know’ in a Saudi learner corpus</td>
</tr>
<tr>
<td>Alzahrani, Najwa</td>
<td>Tracing the meaning of ideological keywords; a diachronic corpus–assisted study of the discursive construction of ‘allibraliah’ in the Saudi context</td>
</tr>
<tr>
<td>Ansarifar, Ahmad</td>
<td>Phrasal Complexity in the Writing of Iranian EFL College–Level Students</td>
</tr>
<tr>
<td>Anthony, Laurence</td>
<td>The use of general academic and discipline–specific corpora for research writing: introducing data–driven learning to PhD students in Hong Kong</td>
</tr>
<tr>
<td></td>
<td>Automatic Creation and Discourse–Level Annotation of Individualized Discipline–Specific Corpora for the Data–Driven Learning (DDL) Classroom</td>
</tr>
<tr>
<td>Aragrande, Gaia</td>
<td>Interpreting web–based populism in Italy: a corpus–based analysis of selected Keywords in Beppe Grillo's blog</td>
</tr>
<tr>
<td>Arppe, Antti</td>
<td>Expanding the coverage of a computational model for an endangered language with a derivational component – the case of Plains Cree</td>
</tr>
<tr>
<td></td>
<td>Corpus Approaches to Northern Haida (*)</td>
</tr>
<tr>
<td>Name</td>
<td>Title</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>----------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Axelsson, Karin</td>
<td>Spoken BNC2014 EAS vs. the demographic part of the BNC: What can a study of tag questions tell us?</td>
</tr>
<tr>
<td>Bai, Yinchun</td>
<td>Computer Assisted Legal Linguistics (CAL²): An interdisciplinary approach</td>
</tr>
<tr>
<td>Baisa, Vit</td>
<td>Simplifying terminology extraction: OneClick Terms</td>
</tr>
<tr>
<td>Baker, Helen</td>
<td>Enriching our understanding of historic drought and water scarcity: investigating 200 years of news texts</td>
</tr>
<tr>
<td>Baker, Paul</td>
<td>Discourse practice and news online: comparing discursive constructions of Romanian immigrants in articles published by the Daily Express with readers' comments</td>
</tr>
<tr>
<td>Banski, Piotr</td>
<td>Legal canvas for a patchwork of multilingual quotations: the case of CoMParS</td>
</tr>
<tr>
<td>Barcellos, Carolina</td>
<td>Conventionality and creativity in translation: a corpus–based case study of translator’s style</td>
</tr>
<tr>
<td>Barlow, Michael</td>
<td>Frequency and sequence: highest, first -- lowest, last</td>
</tr>
<tr>
<td>Barnes, Michael</td>
<td>'We are' and 'they are' constructions and the construction of social categories</td>
</tr>
<tr>
<td>Becker, Israela</td>
<td>The Negation Operator is not a Suppressor of the Concept in its Scope. In Fact, Quite the Opposite</td>
</tr>
<tr>
<td>Berber Sardinha, Tony</td>
<td>Dimensions of collocation in American English — A linguistic typology of American television programs</td>
</tr>
<tr>
<td>Bernardini, Silvia</td>
<td>Words that go together: An exploration of the idiom principle in institutional spoken English</td>
</tr>
<tr>
<td>Betts, Lucy</td>
<td>Understanding street harassment of children: identifying recurrent behaviours in a corpus of young people’s accounts of harassment</td>
</tr>
<tr>
<td>Biber, Douglas</td>
<td>Exploring Methods for Evaluating Corpus Representativeness</td>
</tr>
<tr>
<td>Blything, Liam</td>
<td>Guided reading: Using corpus methods to investigate how teacher strategies differ across children's reading ability, SES, and teacher experience</td>
</tr>
<tr>
<td>Bocorny Finatto, Maria José</td>
<td>Written Proficiency in English for Specific Purposes: a Corpus Study of Abstracts in Health Sciences (*)</td>
</tr>
<tr>
<td>Brenchley, Mark</td>
<td>The Developmental Relationship between Spoken and Written Clause Packaging in an English Secondary School (*)</td>
</tr>
<tr>
<td></td>
<td>--- Growth in Grammar: A multidimensional analysis of student writing between 5 and 16 (*)</td>
</tr>
<tr>
<td>Brezina, Vaclav</td>
<td>How large is the BNC? A proposal for standardised tokenization and word counting</td>
</tr>
<tr>
<td></td>
<td>--- MI-score–based collocations in language learning research: A critical evaluation</td>
</tr>
<tr>
<td></td>
<td>--- Revisiting the idiom principle through the lens of an agglutinating language: A corpus–based description of adjective–noun collocations in Turkish and in English</td>
</tr>
<tr>
<td>Brown, Guy</td>
<td>Metaphorical Constructions in Modern Economic Discourse: A Large-Scale Corpus Analysis</td>
</tr>
<tr>
<td>Bruhova, Gabriela</td>
<td>English presentative semantic patterns – as seen through a parallel translation corpus</td>
</tr>
<tr>
<td>Buckingham, Louisa</td>
<td>The textual colligation of stance phraseology in cross–disciplinary academic discourses: the timing of authors’ self–projection</td>
</tr>
<tr>
<td>Budgell, Brian</td>
<td>&quot;It is important to reinforce the importance...&quot;: ‘Hype’ in medical research articles</td>
</tr>
<tr>
<td>Busse, Beatrix</td>
<td>HeidelGram: Network analysis of grammarians’ references in 19th–century British grammars – a corpus–based study</td>
</tr>
<tr>
<td>Bušta, Jan</td>
<td>JSI Newsfeed corpus</td>
</tr>
<tr>
<td>Name</td>
<td>Title</td>
</tr>
<tr>
<td>-----------------------</td>
<td>--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Cabrera Alvarez, Carmen</td>
<td>Lesis of Spanish Video Game Magazines: Preferred Mechanisms to Adapt Anglicisms (*)</td>
</tr>
<tr>
<td>Cain, Kate</td>
<td>Guided reading: Using corpus methods to investigate how teacher strategies differ across children's reading ability, SES, and teacher experience</td>
</tr>
<tr>
<td>Candarli, Duygu</td>
<td>Cross-cultural and paradigmatic influences on lexical bundles in English academic writing</td>
</tr>
<tr>
<td>Carmody, Emily</td>
<td>Attributing the Bixby Letter using n-gram tracing</td>
</tr>
<tr>
<td>Castello, Erik</td>
<td>A longitudinal corpus-based study of it--extraposition constructions in Italian EFL academic writing</td>
</tr>
<tr>
<td>Čermáková, Anna</td>
<td>From ICE to ICC: A proposal for an International Comparable Corpus --- Corpus-based resources for L1 and L2 teaching of Czech</td>
</tr>
<tr>
<td>Chan, Clarice S. C.</td>
<td>Strategies for incorporating findings from corpus linguistics studies into the teaching of spoken English for business (*)</td>
</tr>
<tr>
<td>Charles, Maggie</td>
<td>Corpus-assisted editing for doctoral students: Do--it--yourself corpora for self--correction and learning</td>
</tr>
<tr>
<td>Chau, Meng Huat</td>
<td>A longitudinal learner corpus study of syntactic complexity in L2 writing</td>
</tr>
<tr>
<td>Chen, Alvin</td>
<td>Assessing Text Difficulty Development in ELT Textbooks Series Using N-gram Language Models based on BNC</td>
</tr>
<tr>
<td>Chen--Hsien, Hao–Jan</td>
<td>Uncovering Informal Expressions Used in English NNS Thesis Writing (*)</td>
</tr>
<tr>
<td>Chen, Howard</td>
<td>Extracting English Verb–Noun Miscollocations from NNS Academic Writing based on Corpus Comparison</td>
</tr>
<tr>
<td>Chen, Meilin</td>
<td>The use of general academic and discipline–specific corpora for research writing: introducing data–driven learning to PhD students in Hong Kong</td>
</tr>
<tr>
<td>Chen, Yaoyao</td>
<td>Multimodal Corpus–based Investigations of the Gesture–Speech Relationship</td>
</tr>
<tr>
<td>Cheng, An Chung</td>
<td>The Development and Evaluation of a Corpus–based Spanish Collocation Error Detection and Revision Suggestion Tool (*)</td>
</tr>
<tr>
<td>Cheng, Yu–Chen</td>
<td>Extracting English Verb–Noun Miscollocations from NNS Academic Writing based on Corpus Comparison</td>
</tr>
<tr>
<td>Chlumská, Lucie</td>
<td>Corpus–based resources for L1 and L2 teaching of Czech</td>
</tr>
<tr>
<td>Choi, Jun</td>
<td>Report on the Automatic Extraction of Korean Scientific Phrasal Term Candidates, with a Focus on Science Textbook Corpus (*)</td>
</tr>
<tr>
<td>Cienki, Alan</td>
<td>Action, metaphor and gesture: A corpus–analytical approach (*)</td>
</tr>
<tr>
<td>Clarke, Ben</td>
<td>Being clear – the politician's riposte?</td>
</tr>
<tr>
<td>Clarke, Isobelle</td>
<td>Dimension of Twitter Trolling: Short Text Classification Using Multiple Correspondence Analysis --- Attributing the Bixby Letter using n–gram tracing</td>
</tr>
<tr>
<td>Coffey, Stephen</td>
<td>From corpora to learners' dictionaries: a case study of the function word 'otherwise'</td>
</tr>
<tr>
<td>Collins, Peter</td>
<td>Exploring Grammatical Colloquialisation in Non–Native English</td>
</tr>
<tr>
<td>Crespo, Begoña</td>
<td>Discipline, discourse and new viewpoints. A Report on the Coruña Corpus (*)</td>
</tr>
<tr>
<td>Culpeper, Jonathan</td>
<td>Grappling with Shakespeare's words: Maximizing historical corpus–based approaches</td>
</tr>
<tr>
<td>Cvrcek, Vaclav</td>
<td>Multi–dimensional analysis of Czech. Pilot study</td>
</tr>
<tr>
<td>Davies, Mark</td>
<td>The new 4 billion word NOW corpus, with 4–5 million words of data added every day</td>
</tr>
<tr>
<td>Davis, Tony</td>
<td>Corpus linguistics and disability studies: Ableism, agency, power, and blindness</td>
</tr>
<tr>
<td>Dayrell, Carmen</td>
<td>Enriching our understanding of historic drought and water scarcity: investigating 200 years of news texts</td>
</tr>
<tr>
<td>Author</td>
<td>Title</td>
</tr>
<tr>
<td>------------------------</td>
<td>---------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>De Felice, Rachele</td>
<td>&quot;Please accept my appreciation&quot;: A corpus–pragmatic investigation of thanking behaviour in British and American emails</td>
</tr>
<tr>
<td>Dong, Jihua</td>
<td>The textual colligation of stance phraseology in cross–disciplinary academic discourses: the timing of authors' self–projection</td>
</tr>
<tr>
<td>Dong, Min</td>
<td>Across discipline corpus–based study of logical relation verb patterns in the research article</td>
</tr>
<tr>
<td>Donnelly, Kevin</td>
<td>Towards a Welsh Semantic Tagger: Creating lexicons for a resource poor language</td>
</tr>
<tr>
<td>Drasovean, Anda</td>
<td>'Do animals have nationality?' Insights from a cross–linguistic corpus–assisted study on animal representation in online newspapers</td>
</tr>
<tr>
<td></td>
<td>Animals and their places in news discourse: insights from cross–linguistic and diachronic perspectives (*)</td>
</tr>
<tr>
<td>Durrant, Philip</td>
<td>Growth in Grammar: A multidimensional analysis of student writing between 5 and 16 (*)</td>
</tr>
<tr>
<td>Dücker, Lisa</td>
<td>Investigating the emergence of noun capitalization in a corpus of handwritten texts</td>
</tr>
<tr>
<td>Dyka, Susanne</td>
<td>A web of analogies: key descriptive constructions in modern fiction (*)</td>
</tr>
<tr>
<td>Ebeling, Jarle</td>
<td>Basic corpus annotation made easy: The Language Analysis Portal (LAP) (*)</td>
</tr>
<tr>
<td>Eder, Maciej</td>
<td>If an atom is a letter, then a molecule is a word: applying corpus linguistic methods to chemistry (*)</td>
</tr>
<tr>
<td>Egbert, Jesse</td>
<td>Fiction—One Register or Two? Narrative and Fictional Speech in Dickens’s Novels</td>
</tr>
<tr>
<td></td>
<td>Exploring Methods for Evaluating Corpus Representativeness</td>
</tr>
<tr>
<td>Evert, Stefan</td>
<td>Making sense of multivariate analyses of linguistic variation (*)</td>
</tr>
<tr>
<td></td>
<td>Reliable measures of syntactic and lexical complexity: The case of Iris Murdoch</td>
</tr>
<tr>
<td>Fankhauser, Peter</td>
<td>Visualizing Language Change in a Corpus of Contemporary German (*)</td>
</tr>
<tr>
<td></td>
<td>Visualization of Corpus Frequencies at Text Level (*)</td>
</tr>
<tr>
<td>Ferraresi, Adriano</td>
<td>Words that go together: An exploration of the idiom principle in institutional spoken English</td>
</tr>
<tr>
<td>Finlayson, Natalie</td>
<td>Corpora, Prototypes and Literary Translation (*)</td>
</tr>
<tr>
<td>Fischer, Stefan</td>
<td>Visualization of Corpus Frequencies at Text Level (*)</td>
</tr>
<tr>
<td>Fitzpatrick, Tess</td>
<td>The CorCenCC Crowdsourcing App: A Bespoke Tool for the User–Driven Creation of the National Corpus of Contemporary Welsh</td>
</tr>
<tr>
<td></td>
<td>&quot;How will you make sure the material is suitable for children?&quot;: User–informed design of Welsh corpus–based learning/teaching tools</td>
</tr>
<tr>
<td>Flowerdew, John</td>
<td>The use of general academic and discipline–specific corpora for research writing: introducing data–driven learning to PhD students in Hong Kong</td>
</tr>
<tr>
<td>Forchini, Pierfranca</td>
<td>The applicability of movies in legal language teaching: Evidence from Multi–Dimensional Analysis</td>
</tr>
<tr>
<td>Fotiadou, Maria</td>
<td>'Competitive job markets' and 'employable graduates': A corpus–based critical discourse analysis of UK universities’ Careers and Employability web pages</td>
</tr>
<tr>
<td>Freitas, Ana Luiza</td>
<td>Written Proficiency in English for Specific Purposes: a Corpus Study of Abstracts in Health Sciences (*)</td>
</tr>
<tr>
<td>Fruttaldo, Antonio</td>
<td>The discursive representation of the US same–sex ruling: A corpus–based investigation of news values in the US, UK and Italian press</td>
</tr>
<tr>
<td>Name</td>
<td>Title</td>
</tr>
<tr>
<td>-----------------------</td>
<td>-----------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Fujimoto, Kazuko</td>
<td>Do English textbooks reflect the actual use of English?: the present perfect and temporal adverbials</td>
</tr>
<tr>
<td>Fuoli, Matteo</td>
<td>Combining corpus and experimental methods to investigate trust-building strategies in corporate discourse</td>
</tr>
<tr>
<td>Gablasova, Dana</td>
<td>MI-score-based collocations in language learning research: A critical evaluation</td>
</tr>
<tr>
<td>Gao, Jie</td>
<td>Variability in citation practices of developing L2 writers in first-year writing courses (*)</td>
</tr>
<tr>
<td>Gardner, Sheena</td>
<td>Stance in the BAWE Corpus: New Revelations from Multidimensional Analysis</td>
</tr>
<tr>
<td>Gather, Kirsten</td>
<td>HeidelGram: Network analysis of grammarians' references in 19th-century British grammars – a corpus-based study</td>
</tr>
<tr>
<td>Gauer, Isabelle</td>
<td>Computer Assisted Legal Linguistics (CAL²): An interdisciplinary approach</td>
</tr>
<tr>
<td>Gee, Matt</td>
<td>“Please please don’t buy this game like I did. I feel terrible and wish I could return it!”: A corpus-based study of professional and consumer reviews of video games</td>
</tr>
<tr>
<td>Gerasimenko, Ekaterina</td>
<td>Design of test-making tools for the learner corpus</td>
</tr>
<tr>
<td>Gerhard, Rampl</td>
<td>A mountain of work. Building an Alpine Heritage Text Corpus (*)</td>
</tr>
<tr>
<td>Gideon, Hannah</td>
<td>Attributing the Bixby Letter using n-gram tracing</td>
</tr>
<tr>
<td>Gonen, Einiat</td>
<td>Imperfect language learning vs. active sound change: The shift [i]&gt;[e] in the verbal pattern hif'il in Modern Hebrew</td>
</tr>
<tr>
<td>Gray, Bethany</td>
<td>Exploring Methods for Evaluating Corpus Representativeness</td>
</tr>
<tr>
<td>Greco, Cristina</td>
<td>The Falange Armata Letters: Authorship Profiling of Linguistic Markers of Style and Ideology in Italian Terrorist Communication</td>
</tr>
<tr>
<td>Gregory, Ian</td>
<td>Using corpora to map language: Geographical Text Analysis of UK poverty</td>
</tr>
<tr>
<td>Gries, Stefan Th.</td>
<td>MERGE: A new recursive approach towards multiword expression extraction and four small validation case studies</td>
</tr>
<tr>
<td>Grieve, Jack</td>
<td>Multivariate analysis of short-term lexical change</td>
</tr>
<tr>
<td>Grieve-Smith, Angus</td>
<td>Attributing the Bixby Letter using n-gram tracing</td>
</tr>
<tr>
<td>Groom, Nicholas</td>
<td>Testing usage-based theories with a representative corpus of nineteenth-century French</td>
</tr>
<tr>
<td>Hadikin, Glenn</td>
<td>KeyWords of success – what words are associated with success in online Citizen Science?</td>
</tr>
<tr>
<td>Hamann, Hanjo</td>
<td>Computer Assisted Legal Linguistics (CAL²): An interdisciplinary approach</td>
</tr>
<tr>
<td>Handford, Michael</td>
<td>Corpus analysis of workplace discourse: The case of the construction industry</td>
</tr>
<tr>
<td>Hardie, Andrew</td>
<td>Guided reading: Using corpus methods to investigate how teacher strategies differ across children’s reading ability, SES, and teacher experience</td>
</tr>
<tr>
<td>Harding, Rachel</td>
<td>Understanding street harassment of children: identifying recurrent behaviours in a corpus of young people’s accounts of harassment</td>
</tr>
<tr>
<td>Hart, Christopher</td>
<td>Combining corpus and experimental methods to investigate trust-building strategies in corporate discourse</td>
</tr>
<tr>
<td>Name</td>
<td>Title</td>
</tr>
<tr>
<td>---------------------------</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>Hartmann, Stefan</td>
<td>Investigating the emergence of noun capitalization in a corpus of</td>
</tr>
<tr>
<td></td>
<td>handwritten texts</td>
</tr>
<tr>
<td></td>
<td>Assessing the diachronic change of a word–formation pattern:</td>
</tr>
<tr>
<td></td>
<td>Frequency, productivity, and interaction patterns</td>
</tr>
<tr>
<td>Hata, Kazuki</td>
<td>Utilising corpus linguistic technologies: quantifications in</td>
</tr>
<tr>
<td></td>
<td>conversation analysis</td>
</tr>
<tr>
<td>Hattori, Kota</td>
<td>Quantifying 'standardness' of the language use in a locality: a study</td>
</tr>
<tr>
<td></td>
<td>with Twitter data</td>
</tr>
<tr>
<td>Hawtin, Abi</td>
<td>The British National Corpus Revisited: Developing Parameters for</td>
</tr>
<tr>
<td></td>
<td>Written BNC2014</td>
</tr>
<tr>
<td>Heffernan, Kevin</td>
<td>Age–related patterns in lexical bundle usage: Evidence from a corpus</td>
</tr>
<tr>
<td></td>
<td>of vernacular Japanese</td>
</tr>
<tr>
<td></td>
<td>Quantifying 'standardness' of the language use in a locality: a study</td>
</tr>
<tr>
<td></td>
<td>with Twitter data</td>
</tr>
<tr>
<td>Heini, Annina</td>
<td>Attributing the Bixby Letter using n–gram tracing</td>
</tr>
<tr>
<td>Hennessey, Anthony</td>
<td>A cookbook of co–occurrence comparison techniques and how they</td>
</tr>
<tr>
<td></td>
<td>relate to the subtleties in your research question</td>
</tr>
<tr>
<td>Herman, Ondřej</td>
<td>JSI Newsfeed corpus</td>
</tr>
<tr>
<td>Horch, Eva</td>
<td>The Fragment Corpus (FrAC) (*)</td>
</tr>
<tr>
<td>Huang, Lan–Fen</td>
<td>Speech rates and unfilled pauses in native and learner speech at the</td>
</tr>
<tr>
<td></td>
<td>CEFR levels C1 and B2</td>
</tr>
<tr>
<td>Huang, Ping–Yu</td>
<td>Academic Collocations in Computer Science Research Articles: A</td>
</tr>
<tr>
<td></td>
<td>Corpus–based Study</td>
</tr>
<tr>
<td>Hughes, Jennifer</td>
<td>The processing of collocations by native and non–native speakers of</td>
</tr>
<tr>
<td></td>
<td>English: Evidence from ERP studies</td>
</tr>
<tr>
<td>Hulden, Mans</td>
<td>Expanding the coverage of a computational model for an endangered</td>
</tr>
<tr>
<td></td>
<td>language with a derivational component – the case of Plains Cree</td>
</tr>
<tr>
<td>Hunt, Sally</td>
<td>Boast and bellow, giggle or chatter: gender and verbs of speech in</td>
</tr>
<tr>
<td></td>
<td>children's fiction</td>
</tr>
<tr>
<td>Ikeo, Reiko</td>
<td>The use of 'will' in present tense narrative</td>
</tr>
<tr>
<td>Issitt, Steve</td>
<td>Assessing the impact of an intensive EAP programme:</td>
</tr>
<tr>
<td></td>
<td>multidimensional analysis of an authentic learner corpus</td>
</tr>
<tr>
<td>Izquierdo Alegría, Dámaso</td>
<td>Is a CADS approach useful to empirically measure populism? A</td>
</tr>
<tr>
<td></td>
<td>comparative analysis of the representation of the French &quot;peuple&quot;</td>
</tr>
<tr>
<td></td>
<td>and their threats in the discourse of the major political parties in</td>
</tr>
<tr>
<td></td>
<td>France</td>
</tr>
<tr>
<td>Jakubicek, Milos</td>
<td>JSI Newsfeed corpus</td>
</tr>
<tr>
<td>Jeaco, Stephen</td>
<td>&quot;What kind of neighbourhood is that?&quot;: Helping learners explore the</td>
</tr>
<tr>
<td></td>
<td>semantic neighbourhood of words and phrases</td>
</tr>
<tr>
<td>Jenset, Gard</td>
<td>Keeping the English dative alternation in the family: a quantitative</td>
</tr>
<tr>
<td></td>
<td>corpus–based study of spoken data</td>
</tr>
<tr>
<td>Ji, Jie</td>
<td>Animacy hierarchy within inanimate nouns: English corpus evidence</td>
</tr>
<tr>
<td></td>
<td>from a prototypical perspective</td>
</tr>
<tr>
<td>Jin, Bixi</td>
<td>Using Multidimensional Analysis to Investigate the RA Discussion</td>
</tr>
<tr>
<td></td>
<td>Sections of Journal Articles in Chemical Engineering</td>
</tr>
<tr>
<td>Joulain–Jay, Amelia</td>
<td>Grappling with Shakespeare's words: Maximizing historical corpus–</td>
</tr>
<tr>
<td></td>
<td>based approaches</td>
</tr>
<tr>
<td>Jung, Hae–Yun</td>
<td>Report on the Automatic Extraction of Korean Scientific Phrasal Term</td>
</tr>
<tr>
<td></td>
<td>Candidates, with a Focus on Science Textbook Corpus (*)</td>
</tr>
<tr>
<td>Kamocki, Paweł</td>
<td>Legal canvas for a patchwork of multilingual quotations: the case of</td>
</tr>
<tr>
<td></td>
<td>CoMParS</td>
</tr>
<tr>
<td>Kang, Hyeonah</td>
<td>Report on the Automatic Extraction of Korean Scientific Phrasal Term</td>
</tr>
<tr>
<td></td>
<td>Candidates, with a Focus on Science Textbook Corpus (*)</td>
</tr>
<tr>
<td>Kaufmann, Kean</td>
<td>Corpus linguistics and disability studies: Ableism, agency, power,</td>
</tr>
<tr>
<td></td>
<td>and blindness</td>
</tr>
<tr>
<td>Author</td>
<td>Title</td>
</tr>
<tr>
<td>-------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Kehoe, Andrew</td>
<td>“Please please don't buy this game like I did. I feel terrible and wish I could return it!”: A corpus–based study of professional and consumer reviews of video games</td>
</tr>
<tr>
<td>Kemp, Jenny</td>
<td>The building blocks of a Discipline-Specific Vocabulary Core (DSVC) (*)</td>
</tr>
<tr>
<td>Keng, Nicole</td>
<td>Corpus tools for an Academic Writing course: a case study (*)</td>
</tr>
<tr>
<td>Kermes, Hannah</td>
<td>Average surprisal of parts-of-speech</td>
</tr>
<tr>
<td>Kim, Seiyeon</td>
<td>Report on the Automatic Extraction of Korean Scientific Phrasal Term Candidates, with a Focus on Science Textbook Corpus (*)</td>
</tr>
<tr>
<td>Kirk, John M.</td>
<td>From ICE to ICC: A proposal for an International Comparable Corpus</td>
</tr>
<tr>
<td>Kirsimäe, Merli</td>
<td>Lexicogrammatical features in the Estonian spoken mini–corpus of English as a lingua franca (*)</td>
</tr>
<tr>
<td>Kishie, Shunsuke</td>
<td>Quantifying 'standardness' of the language use in a locality: a study with Twitter data</td>
</tr>
<tr>
<td>Kiš Žuvela, Sanja</td>
<td>The conceptualization of music in semantic frames based on word sketches (*)</td>
</tr>
<tr>
<td>Klavan, Jane</td>
<td>Lexicogrammatical features in the Estonian spoken mini–corpus of English as a lingua franca (*)</td>
</tr>
<tr>
<td>Klimová, Jana</td>
<td>NomVallex: Valency Patterns of Semantically Classified Czech Nouns</td>
</tr>
<tr>
<td>Knight, Dawn</td>
<td>The CorCenCC Crowdsourcing App: A Bespoke Tool for the User-Driven Creation of the National Corpus of Contemporary Welsh ---</td>
</tr>
<tr>
<td></td>
<td>&quot;How will you make sure the material is suitable for children?&quot;: User-informed design of Welsh corpus–based learning/teaching tools</td>
</tr>
<tr>
<td></td>
<td>Towards a Welsh Semantic Tagger: Creating Lexicons for A Resource Poor Language</td>
</tr>
<tr>
<td></td>
<td>Creating a Bespoke Corpus Sampling Frame for a Minoritised Language: CorCenCC, the National Corpus of Contemporary Welsh</td>
</tr>
<tr>
<td>Kobayashi, Yuichiro</td>
<td>Developmental Patterns of Metadiscourse in Second Language Writing (*)</td>
</tr>
<tr>
<td>Kolářová, Veronika</td>
<td>NomVallex: Valency Patterns of Semantically Classified Czech Nouns</td>
</tr>
<tr>
<td>Komrsková, Zuzana</td>
<td>Multi-dimensional analysis of Czech. Pilot study</td>
</tr>
<tr>
<td>Kouyilekov, Milen</td>
<td>Basic corpus annotation made easy: The Language Analysis Portal (LAP)</td>
</tr>
<tr>
<td>Kováříková, Dominika</td>
<td>Exploring Corpus Data Using Data Mining: A Project of Automatic Term Recognition</td>
</tr>
<tr>
<td>Kreischer, Kim-Sue</td>
<td>Intertextuality and ideology in discourse</td>
</tr>
<tr>
<td>Krek, Simon</td>
<td>JSI Newsfeed corpus</td>
</tr>
<tr>
<td>Kryzhanivska, Anastasiia</td>
<td>Do You Also See what I See: Russian-Ukrainian Conflict in European, Ukrainian and Russian Media ---</td>
</tr>
<tr>
<td></td>
<td>Corpus linguistics and disability studies: Ableism, agency, power, and blindness</td>
</tr>
<tr>
<td>Kunilovskaya, Maria</td>
<td>Frequency distribution and usage of discourse markers in English to Russian translation: learner vs. professional translatinal behavior</td>
</tr>
<tr>
<td>Kupietz, Marc</td>
<td>Visualizing Language Change in a Corpus of Contemporary German (*)</td>
</tr>
<tr>
<td>Lachler, Jordan</td>
<td>Corpus Approaches to Northern Haida (*)</td>
</tr>
<tr>
<td>Lan, Ge</td>
<td>Variability in citation practices of developing L2 writers in first-year writing courses (*)</td>
</tr>
<tr>
<td>Authors</td>
<td>Title</td>
</tr>
<tr>
<td>-------------------------</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>Lapponi, Emanuele</td>
<td>Basic corpus annotation made easy: The Language Analysis Portal (LAP) (*)</td>
</tr>
<tr>
<td>Laranetto, Helena</td>
<td>Metaphorical Constructions in Modern Economic Discourse: A Large-Scale Corpus Analysis</td>
</tr>
<tr>
<td>Larl, Bettina</td>
<td>Geolocating German on Twitter – Hitches and Glitches of Building and Exploring a Twitter Corpus (*)</td>
</tr>
<tr>
<td>Lastres-López, Cristina</td>
<td>The pragmatics of insubordinate if-clauses in British English: A corpus-based study</td>
</tr>
<tr>
<td>Le Foll, Elen</td>
<td>Textbook English: A corpus-based analysis of language use in German and French EFL textbooks (*)</td>
</tr>
<tr>
<td>Lederer, Jenny</td>
<td>Metaphorical Constructions in Modern Economic Discourse: A Large-Scale Corpus Analysis</td>
</tr>
<tr>
<td>Levshina, Natalia</td>
<td>Bringing together corpus linguistics and typology: Frequency, informativity and grammatical asymmetries</td>
</tr>
<tr>
<td>Li, Long</td>
<td>A Corpus-based Approach to the Patterns of Motivated and Ideologically-invested Shifts in Translation</td>
</tr>
<tr>
<td>Li, Xi</td>
<td>A Corpus-based Approach to the Patterns of Motivated and Ideologically-invested Shifts in Translation</td>
</tr>
<tr>
<td>Liang, Maocheng</td>
<td>Animacy hierarchy within inanimate nouns: English corpus evidence from a prototypical perspective</td>
</tr>
<tr>
<td>Liesenfeld, Andreas</td>
<td>Extracting construction networks from Cantonese speech corpora using clustering algorithms (*)</td>
</tr>
<tr>
<td>Lin, Fuying</td>
<td>A Case Study of the Characterizations of Spoken Academic English: Adjective–Noun Combination</td>
</tr>
<tr>
<td>Lin, Ling</td>
<td>The textual colligation of stance phraseology in cross-disciplinary academic discourses: the timing of authors’ self-projection --- Lexical bundles in academic bio-data: A corpus interdisciplinary study</td>
</tr>
<tr>
<td>Lin, Phoebe</td>
<td>A new tool for concordancing the Web as a multimodal corpus (*)</td>
</tr>
<tr>
<td>Lischinsky, Alon</td>
<td>Distant reading intimate encounters: a big data approach to online erotica</td>
</tr>
<tr>
<td>Liu, Tanjun</td>
<td>Evaluating the effect of data-driven learning (DDL) on the acquisition of academic collocations by advanced Chinese learners of English (*)</td>
</tr>
<tr>
<td>Liu, Yang</td>
<td>Verb–Argument Constructions in Chinese EFL Learner’s Oral Production</td>
</tr>
<tr>
<td>Love, Robbie</td>
<td>Bad language revisited: swearing in the Spoken BNC2014</td>
</tr>
<tr>
<td>Lu, Hui–Chuan</td>
<td>The Development and Evaluation of a Corpus-based Spanish Collocation Error Detection and Revision Suggestion Tool (*)</td>
</tr>
<tr>
<td>Lugli, Ligeia</td>
<td>Lexis and Tradition: Variation in the vocabulary of Buddhist Literature (*)</td>
</tr>
<tr>
<td>Lukac, Morana</td>
<td>A corpus-based approach to investigating twenty-first-century prescriptivism</td>
</tr>
<tr>
<td>Lukeš, David</td>
<td>Multi-dimensional analysis of Czech. Pilot study</td>
</tr>
<tr>
<td>Macdonald, Lindsey</td>
<td>Variability in citation practices of developing L2 writers in first-year writing courses (*)</td>
</tr>
<tr>
<td>Machálek, Tomáš</td>
<td>KonText – a modern, customizable corpus query interface</td>
</tr>
<tr>
<td>Mahlberg, Michaela</td>
<td>Fiction—One Register or Two? Narrative and Fictional Speech in Dickens’s Novels --- A cookbook of co-occurrence comparison techniques and how they relate to the subtleties in your research question</td>
</tr>
</tbody>
</table>
Mahmood, Khalid  
The Ideological Representation of ‘self’ and ‘other’ in Post 9/11 Discourse in British Newspapers

Mala, Marketa  
English presentative semantic patterns – as seen through a parallel translation corpus

Malamatidou, Sofia  
Corpus triangulation: Towards a new methodological framework for translation studies

Mark, Geraldine  
English Grammar Profile: a corpus-based inventory of English grammar competency using the Cambridge Learner Corpus

Marshall, Lindsay  
The CorCenCC Crowdsourcing App: A Bespoke Tool for the User-Driven Creation of the National Corpus of Contemporary Welsh

Mastropierro, Lorenzo  
Using key words as nodes of cohesive networks

Matuška, Ondřej  
Simplifying terminology extraction: OneClick Terms

Mautner, Gerlinde  
Rigour, Relevance, Reflection: CL Methodology Through a Critical Lens

McCloughlin, Emma  
Animals and their places in news discourse: insights from cross-linguistic and diachronic perspectives (*)
---
Badgers, hedgehogs and squirrels: a diachronic corpus-assisted discourse study of British wildlife in the news

McDonald, Daniel  
Changing Discourse-Semantics of Risk – A historical analysis of US newspapers after World War II utilizing corpus linguistics research instruments

Mcenery, Tony  
Enriching our understanding of historic drought and water scarcity: investigating 200 years of news texts
---
MI-score-based collocations in language learning research: A critical evaluation

McGillivray, Barbara  
Keeping the English dative alternation in the family: a quantitative corpus-based study of spoken data

McGlashan, Mark  
Discourse practice and news online: comparing discursive constructions of Romanian immigrants in articles published by the Daily Express with readers’ comments

Meiriño-Gómez, Jesús  
Designing, Compiling and Sampling Specialised Corpora for Researching Metaphor in Translation: The Case of SCoPE (*)

Menke, Peter  
An annotated video corpus of interactions dealing with the collaborative construction of fiction in games (*)

Michelfeit, Jan  
Simplifying terminology extraction: OneClick Terms

Miličević, Maja  
Words that go together: An exploration of the idiom principle in institutional spoken English

Millar, Neil  
"It is important to reinforce the importance…": ‘Hype’ in medical research articles

Modrzyk, Urszula  
If an atom is a letter, then a molecule is a word: applying corpus linguistic methods to chemistry (*)

Monaco, Leida Maria  
Discipline, discourse and new viewpoints. A Report on the Coruña Corpus (*)

Morris, Steve  
The CorCenCC Crowdsourcing App: A Bespoke Tool for the User-Driven Creation of the National Corpus of Contemporary Welsh
---
"How will you make sure the material is suitable for children?": User-informed design of Welsh corpus-based learning/teaching tools
---
Creating a Bespoke Corpus Sampling Frame for a Minoritised Language: CorCenCC, the National Corpus of Contemporary Welsh

Morton, Ralph  
Public Archives as a Source of Historical Linguistic Data: The Construction and Analysis of the British Telecom Correspondence Corpus
<table>
<thead>
<tr>
<th>Author</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moskowich, Isabel</td>
<td>Discipline, discourse and new viewpoints. A Report on the Coruña Corpus (*)</td>
</tr>
<tr>
<td>Muguiro, Natalia</td>
<td>Citing external sources in Educational Neuroscience articles: in search of an interdisciplinary stance and voice</td>
</tr>
<tr>
<td>Murphy, Amanda</td>
<td>Investigating L2 errors in a quasi–longitudinal learner English corpus, with particular reference to word order and the position of also</td>
</tr>
<tr>
<td>Murphy, M. Lynne</td>
<td>“Please accept my appreciation”: A corpus–pragmatic investigation of thanking behaviour in British and American emails</td>
</tr>
<tr>
<td>Mwinaaru, Isaac</td>
<td>Lexical bundles in academic bio–data: A corpus interdisciplinary study</td>
</tr>
<tr>
<td>Nacey, Susan</td>
<td>Metaphorical analogies in an online discussion forum for relationship abuse survivors: “That analogy really works for me”</td>
</tr>
<tr>
<td>Nardone, Chiara</td>
<td>Interpreting web–based populism in Italy: a corpus–based analysis of selected Keywords in Beppe Grillo's blog</td>
</tr>
<tr>
<td>Naser Eddine, Abeer</td>
<td>A corpus–based analysis of educational curricula in the Arab States against European curricula and UNESCO benchmarks: Lexical networks and cognitive function of ICT terminology</td>
</tr>
<tr>
<td>Nasseri, Maryam</td>
<td>A Corpus–based Analysis of Syntactic Complexity measures in the Academic Writing of EFL, ESL, and Native English Master’s Students</td>
</tr>
<tr>
<td>Neale, Steven</td>
<td>The CorCenCC Crowdsourcing App: A Bespoke Tool for the User–Driven Creation of the National Corpus of Contemporary Welsh --- &quot;How will you make sure the material is suitable for children?&quot;: User–informed design of Welsh corpus–based learning/teaching tools --- Creating a Bespoke Corpus Sampling Frame for a Minoritised Language: CorCenCC, the National Corpus of Contemporary Welsh</td>
</tr>
<tr>
<td>Needs, Jennifer</td>
<td>The CorCenCC Crowdsourcing App: A Bespoke Tool for the User–Driven Creation of the National Corpus of Contemporary Welsh --- &quot;How will you make sure the material is suitable for children?&quot;: User–informed design of Welsh corpus–based learning/teaching tools --- Creating a Bespoke Corpus Sampling Frame for a Minoritised Language: CorCenCC, the National Corpus of Contemporary Welsh</td>
</tr>
<tr>
<td>Nesi, Hilary</td>
<td>Stance in the BAWE Corpus: New Revelations from Multidimensional Analysis --- An analysis of stance and voice in Applied Linguistics research articles across Mainland Chinese and British cultures</td>
</tr>
<tr>
<td>Newbold, Kendall</td>
<td>Understanding street harassment of children: identifying recurrent behaviours in a corpus of young people’s accounts of harassment</td>
</tr>
<tr>
<td>Nini, Andrea</td>
<td>Attributing the Bixby Letter using n–gram tracing</td>
</tr>
<tr>
<td>Novak, Blaž</td>
<td>JSI Newsfeed corpus</td>
</tr>
<tr>
<td>Ni Loingsigh, Katie</td>
<td>The Design and Development of Corpas na Gaeilge Comhaimseartha (Corpus of Contemporary Irish)</td>
</tr>
<tr>
<td>Nöth, Elmar</td>
<td>Reliable measures of syntactic and lexical complexity: The case of Iris Murdoch</td>
</tr>
<tr>
<td>Ó Cleircin, Gearóid</td>
<td>The Design and Development of Corpas na Gaeilge Comhaimseartha (Corpus of Contemporary Irish)</td>
</tr>
<tr>
<td>Ó Raghallaigh, Brian</td>
<td>The Design and Development of Corpas na Gaeilge Comhaimseartha (Corpus of Contemporary Irish)</td>
</tr>
<tr>
<td>O'Donnell, Mick</td>
<td>The role of contexts of use in the study of learner language</td>
</tr>
<tr>
<td>O'Keeffe, Anne</td>
<td>English Grammar Profile: a corpus–based inventory of English grammar competency using the Cambridge Learner Corpus</td>
</tr>
<tr>
<td>Okada, Sadayuki</td>
<td>The Asymmetry of Conceptual Expansions in Predicational and Modificational Contexts (*)</td>
</tr>
<tr>
<td>Oksuz, Dogus</td>
<td>Revisiting the idiom principle through the lens of an agglutinating language: A corpus–based description of adjective–noun collocations in Turkish and in English</td>
</tr>
<tr>
<td>Name</td>
<td>Title</td>
</tr>
<tr>
<td>------</td>
<td>-------</td>
</tr>
<tr>
<td>Olagboyega, Kolawole</td>
<td>&quot;Japanese English&quot;: A Descriptive Grammar of Educated Written English in Japan</td>
</tr>
<tr>
<td>Omidian, Taha</td>
<td>Situating Lexical Bundles in the Moves and Steps of Applied Linguistics Research Article Introductions</td>
</tr>
<tr>
<td>Onder–Ozdemir, Neslihan</td>
<td>Exploring the interaction across the texts of non–native speakers of English: Compiling reliable learner corpora to identify hedging and boosting in argumentative writing in EFL context</td>
</tr>
<tr>
<td>Orpin, Deborah</td>
<td>‘Never been proven to work in the real world’: appeals to the notion of ‘the real world’ as a discursive strategy in vaccine–critical discourse</td>
</tr>
<tr>
<td>Ostroški Anić, Ana</td>
<td>The conceptualization of music in semantic frames based on word sketches (*)</td>
</tr>
<tr>
<td>Oviedo, Alejandro</td>
<td>El Corpus LESCO. A database for the description of the Costa Rican Sign Language (*)</td>
</tr>
<tr>
<td>Pace–Sigge, Michael</td>
<td>Can Lexical Priming be detected in conversation turn–taking strategies? --- When is a metaphor not a metaphor? – An investigation into lexical characteristics of metaphoricity amongst uncertain cases</td>
</tr>
<tr>
<td>Paterson, Laura</td>
<td>Using corpora to map language: Geographical Text Analysis of UK poverty</td>
</tr>
<tr>
<td>Patten, Amanda</td>
<td>Searching for the unsearchable: A study of English NP Inversion</td>
</tr>
<tr>
<td>Patterson, Katie</td>
<td>When is a metaphor not a metaphor? – An investigation into lexical characteristics of metaphoricity amongst uncertain cases</td>
</tr>
<tr>
<td>Peart, Sheine</td>
<td>Understanding street harassment of children: identifying recurrent behaviours in a corpus of young people’s accounts of harassment</td>
</tr>
<tr>
<td>Pezik, Piotr</td>
<td>Prosody–based clustering of spoken corpus search results (*)</td>
</tr>
<tr>
<td>Philip, Gill</td>
<td>The idiom principle and translated texts</td>
</tr>
<tr>
<td>Piao, Scott</td>
<td>Towards a Welsh Semantic Tagger: Creating Lexicons for A Resource Poor Language</td>
</tr>
<tr>
<td>Pihlaja, Stephen</td>
<td>‘Once I stopped believing in Santa Claus…’: Doing Corpus–Assisted Discourse Analysis on YouTube and Facebook Religious Talk</td>
</tr>
<tr>
<td>Piperski, Alexander</td>
<td>Sum of Minimum Frequencies as a Measure of Corpus Similarity</td>
</tr>
<tr>
<td>Pishghadam, Reza</td>
<td>Phrasal Complexity in the Writing of Iranian EFL College–Level Students</td>
</tr>
<tr>
<td>Plappert, Garry</td>
<td>Candidate Knowledge? Exploring epistemic claims in scientific writing: A corpus–driven approach</td>
</tr>
<tr>
<td>Plevoets, Koen</td>
<td>Using the CORE Corpus and a multivariate analysis to (re)examine the impact of register and structural factors on that/zero complementizer variation in five mental state verbs (MSVs)</td>
</tr>
<tr>
<td>Poli, Francesca</td>
<td>Investigating L2 errors in a quasi–longitudinal learner English corpus, with particular reference to word order and the position of also</td>
</tr>
<tr>
<td>Pollak, Senja</td>
<td>Corpus–assisted discourse analysis of family legislation changes in Slovene media</td>
</tr>
<tr>
<td>Popoola, Olumide</td>
<td>A dictionary, a survey and a corpus walked into a courtroom...: An evaluation of resources for adjudicating meaning in trademark disputes</td>
</tr>
<tr>
<td>Posch, Claudia</td>
<td>“Homo Austriacus” on the mountains. The discursive construction of mountaineering in an alpine heritage corpus --- A mountain of work. Building an Alpine Heritage Text Corpus (*)</td>
</tr>
<tr>
<td>Potts, Amanda</td>
<td>Murderer, mother, slave, or skivvy: XML annotation to enable social actor analysis in a small corpus of English sentencing remarks for women who kill</td>
</tr>
<tr>
<td>Poukarová, Petra</td>
<td>Multi–dimensional analysis of Czech. Pilot study</td>
</tr>
<tr>
<td>Author</td>
<td>Title</td>
</tr>
<tr>
<td>------------------------</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>Powell, Emily</td>
<td>‘I will force myself to believe that everyone is just another monster from doom’: A corpus stylistic analysis of agency in pre-massacre narratives</td>
</tr>
<tr>
<td>Puente–Castelo, Luis</td>
<td>Discipline, discourse and new viewpoints. A Report on the Coruña Corpus (*)</td>
</tr>
<tr>
<td></td>
<td>Conditionals in Eighteenth and Nineteenth century scientific writing</td>
</tr>
<tr>
<td>Qian, Li</td>
<td>Use of Modal Verbs in English Writing by EFL Learners</td>
</tr>
<tr>
<td>Qiu, Zhuang</td>
<td>Capturing Poe in the Gothic Tales —— Exploring Allan Poe’s Stylistic Distinctiveness from a Computational Stylistic Approach</td>
</tr>
<tr>
<td>Raušová, Veronika</td>
<td>Hedging in academic discourse: linguistic research articles (*)</td>
</tr>
<tr>
<td>Rayson, Paul</td>
<td>Towards a Welsh Semantic Tagger: Creating Lexicons for A Resource Poor Language</td>
</tr>
<tr>
<td>Rees, Mair</td>
<td>Creating a Bespoke Corpus Sampling Frame for a Minoritised Language: CorCenCC, the National Corpus of Contemporary Welsh</td>
</tr>
<tr>
<td>Řehořková, Anna</td>
<td>Multi-dimensional analysis of Czech. Pilot study</td>
</tr>
<tr>
<td>Reich, Ingo</td>
<td>The Fragment Corpus (FrAC) (*)</td>
</tr>
<tr>
<td>Reichelt, Susan</td>
<td>Adapting the BNC for sociolinguistic research – a case study on negative concord</td>
</tr>
<tr>
<td>Reshef, Yael</td>
<td>Imperfect language learning vs. active sound change: The shift [i]&gt;[e] in the verbal pattern hif’il in Modern Hebrew</td>
</tr>
<tr>
<td>Rosa, Rodrigo</td>
<td>Strategies of subject indetermination in English and in Brazilian Portuguese (*)</td>
</tr>
<tr>
<td>Ruano, Pablo</td>
<td>Direct Thought Presentation in Charles Dickens’s Fifteen Novels: A Corpus–Stylistic Approach</td>
</tr>
<tr>
<td>Rundell, Michael</td>
<td>Keeping the English dative alternation in the family: a quantitative corpus–based study of spoken data</td>
</tr>
<tr>
<td>Saiful Bahri, Adam</td>
<td>A longitudinal learner corpus study of syntactic complexity in L2 writing</td>
</tr>
<tr>
<td>Ramadhan</td>
<td></td>
</tr>
<tr>
<td>Salager Meyer, Francoise</td>
<td>“It is important to reinforce the importance...”: ‘Hype’ in medical research articles</td>
</tr>
<tr>
<td>Santos, Giovanni</td>
<td>‘This is sort of what you asked, you know’: A comparative corpus–based analysis of pragmatic markers acquired in a study abroad context (*)</td>
</tr>
<tr>
<td>Sato, Yo</td>
<td>Age–related patterns in lexical bundle usage: Evidence from a corpus of vernacular Japanese</td>
</tr>
<tr>
<td></td>
<td>Quantifying 'standardness' of the language use in a locality: a study with Twitter data</td>
</tr>
<tr>
<td>Schmirler, Katherine</td>
<td>Expanding the coverage of a computational model for an endangered language with a derivational component – the case of Plains Cree</td>
</tr>
<tr>
<td>Schneider, Gerold</td>
<td>Saying Whatever It Takes: Creating and Analyzing Corpora from US Presidential Debate Transcripts</td>
</tr>
<tr>
<td>Sepehri, Mehrdad</td>
<td>How Data–Driven Learning Can Help EFL Learners Improve Their Micro Level Skills of Writing</td>
</tr>
<tr>
<td>Shahriari, Hesamoddin</td>
<td>Phrasal Complexity in the Writing of Iranian EFL College–Level Students</td>
</tr>
<tr>
<td></td>
<td>Situating Lexical Bundles in the Moves and Steps of Applied Linguistics Research Article Introductions</td>
</tr>
<tr>
<td>Shank, Christopher</td>
<td>Using the CORE Corpus and a multivariate analysis to (re)examine the impact of register and structural factors on that/zero complementizer variation in five mental state verbs (MSVs)</td>
</tr>
<tr>
<td>Shimazumi, Marilisa</td>
<td>Investigating mentoring in a teacher development programme from a corpus–based hermeneutical–phenomenological perspective</td>
</tr>
<tr>
<td>Siepmann, Dirk</td>
<td>A web of analogies: key descriptive constructions in modern fiction (*)</td>
</tr>
</tbody>
</table>
### Dictionaries and spoken language: the beginnings of a second corpus revolution?

---

* Tolkien as a stylist: a corpus–based investigation into "The Lord of the Rings"  

**Silfverberg, Miikka**  
Expanding the coverage of a computational model for an endangered language with a derivational component – the case of Plains Cree

**Sijanova-Chanturia, Anna**  
Longitudinal investigation of vocabulary development in learner writing

**Sjolin Knight, Catarina**  
Understanding street harassment of children: identifying recurrent behaviours in a corpus of young people’s accounts of harassment

**Skrabal, Michal**  
A corpus of contemporary Czech poetry

**Sliashynskaya, Hanna**  
Multilingualism in Chile: representations of languages in Chilean news media discourse

**Smith, Gillian**  
Using corpus methods to investigate teacher scaffolding in Special Education Needs (SEN) classrooms

**Smith, Simon**  
SkELL: A Discovery–Based Chinese Learning Platform (*)

**Sobočan, Ana Marija**  
Corpus-assisted discourse analysis of family legislation changes in Slovene media

**Song, Lijuan**  
Effects of L1 transfer on L2 learners' VN collocational use: A corpus–based study from semantic preference and semantic prosody perspective (*)

**Sorg, Abberley**  
Corpus linguistics and disability studies: Ableism, agency, power, and blindness

**Sotelo, Patricia**  
Translation–oriented annotation of a multimedia parallel corpus of subtitles (*)

**Spasić, Irena**  
The CorCenCC Crowdsourcing App: A Bespoke Tool for the User–Driven Creation of the National Corpus of Contemporary Welsh

**Spina, Stefania**  
Longitudinal investigation of vocabulary development in learner writing

**Staples, Shelley**  
Variability in citation practices of developing L2 writers in first–year writing courses (*)

**Su, Hang**  
"I’m really sorry that I disappointed you": A local grammar of apology

**Su, Qi**  
Capturing Poe in the Gothic Tales –– Exploring Allan Poe’s Stylistic Distinctiveness from a Computational Stylistic Approach

**Suen, Amy O Y**  
Mining a corpus of online hotel reviews: a pilot study (*)

**Szczepaniak, Renata**  
Investigating the emergence of noun capitalization in a corpus of handwritten texts

**Tagnin, Stella E. O.**  
A bidirectional English–Portuguese Dictionary of Verbal Collocations

**Tanaka, Hiroya**  
Developing a TV drama corpus–informed general spoken formulas list for elementary–level EFL learners (*)

**Teich, Elke**  
Visualization of Corpus Frequencies at Text Level (*)  
---  
Average surprisal of parts–of–speech

**Tench, Christopher R.**  
A cookbook of co–occurrence comparison techniques and how they relate to the subtleties in your research question

**Thomas, Enlli**  
"How will you make sure the material is suitable for children?": User–informed design of Welsh corpus–based learning/teaching tools

**Thompson, Paul**  
Automatic Creation and Discourse–Level Annotation of Individualized Discipline–Specific Corpora for the Data–Driven Learning (DDL) Classroom

**Timperley, Matt**  
How large is the BNC? A proposal for standardised tokenization and word counting

**Tong, Yao**  
Action, metaphor and gesture: A corpus–analytical approach (*)

**Trawinski, Beata**  
Legal canvas for a patchwork of multilingual quotations: the case of CoMParS
<table>
<thead>
<tr>
<th>Author</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vaičenonienė, Jurgita</td>
<td>A Comparable Corpus of Original and Translated Lithuanian: Design and Preliminary Findings (*)</td>
</tr>
<tr>
<td>Vališová, Pavlína</td>
<td>Corpus–based resources for L1 and L2 teaching of Czech</td>
</tr>
<tr>
<td>Valvasion, Elena</td>
<td>Mixed Voices: A Corpus–based Study of Hungarian Parliamentary Debates about the European Union</td>
</tr>
<tr>
<td>Veirano Pinto, Marcia</td>
<td>A linguistic typology of American television programs</td>
</tr>
<tr>
<td>Venuti, Marco</td>
<td>The discursive representation of the US same–sex ruling: A corpus–based investigation of news values in the US, UK and Italian press</td>
</tr>
<tr>
<td>Vernerová, Anna</td>
<td>NomVallex: Valency Patterns of Semantically Classified Czech Nouns</td>
</tr>
<tr>
<td>Vida, Balázs</td>
<td>Corpora in teacher training through Pedagogical Grammar – A case study of two workshops</td>
</tr>
<tr>
<td>Viggiano, Claudia</td>
<td>Community–specific language in online citizen science forums: a corpus–driven diachronic study</td>
</tr>
<tr>
<td>Vinogradova, Olga</td>
<td>Design of test–making tools for the learner corpus</td>
</tr>
<tr>
<td>Vogel, Friedemann</td>
<td>Computer Assisted Legal Linguistics (CAL²): An interdisciplinary approach</td>
</tr>
<tr>
<td>Vrana, Leo</td>
<td>Saying Whatever It Takes: Creating and Analyzing Corpora from US Presidential Debate Transcripts</td>
</tr>
<tr>
<td>Wahl, Alexander</td>
<td>MERGE: A new recursive approach towards multiword expression extraction and four small validation case studies</td>
</tr>
<tr>
<td>Waibel, Emily</td>
<td>Attributing the Bixby Letter using n–gram tracing</td>
</tr>
<tr>
<td>Wang, Bingxin</td>
<td>The evaluative know–how of Chinese scholars: A contrastive corpus analysis of evaluative it patterns in research articles</td>
</tr>
<tr>
<td>Wang, Hua</td>
<td>Exploring the Integration of Valency Patterns and Meaning</td>
</tr>
<tr>
<td>Wang, Ying</td>
<td>Formulaicity across academic disciplines: a function–to–form approach</td>
</tr>
<tr>
<td>Wang, Yingying</td>
<td>A Parallel Corpus–based Study of Chinese Arabic verb phrase alignment</td>
</tr>
<tr>
<td>Wang, Zhaozhe</td>
<td>Variability in citation practices of developing L2 writers in first–year writing courses (*)</td>
</tr>
<tr>
<td></td>
<td>A Corpus Approach to the Disciplinary History of Second Language Writing (*)</td>
</tr>
<tr>
<td>Wankerl, Sebastian</td>
<td>Reliable measures of syntactic and lexical complexity: The case of Iris Murdoch</td>
</tr>
<tr>
<td>Watkins, Gareth</td>
<td>The CorCenCC Crowdsourcing App: A Bespoke Tool for the User–Driven Creation of the National Corpus of Contemporary Welsh</td>
</tr>
<tr>
<td></td>
<td>Towards a Welsh Semantic Tagger: Creating Lexicons for A Resource Poor Language</td>
</tr>
<tr>
<td></td>
<td>Creating a Bespoke Corpus Sampling Frame for a Minoritised Language: CorCenCC, the National Corpus of Contemporary Welsh</td>
</tr>
<tr>
<td>Wei, Naixing</td>
<td>The evaluative know–how of Chinese scholars: A contrastive corpus analysis of evaluative it patterns in research articles</td>
</tr>
<tr>
<td>Weisser, Martin</td>
<td>Doing Corpus Pragmatics in DART 2.0 – New and Improved Ways</td>
</tr>
<tr>
<td>Wells–Jensen, Jason</td>
<td>Corpus linguistics and disability studies: Ableism, agency, power, and blindness</td>
</tr>
<tr>
<td>Wells–Jensen, Sheri</td>
<td>Corpus linguistics and disability studies: Ableism, agency, power, and blindness</td>
</tr>
<tr>
<td>Whiteside, Karin</td>
<td>The semantic patterning of grammatical keywords in undergraduate History and PIR (Politics &amp; international Relations) essays: a corpus–driven investigation (*)</td>
</tr>
<tr>
<td>Wiegand, Viola</td>
<td>A cookbook of co–occurrence comparison techniques and how they relate to the subtleties in your research question</td>
</tr>
<tr>
<td>Name</td>
<td>Title</td>
</tr>
<tr>
<td>---------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Winter, Bodo</td>
<td>English taste and smell adjectives as generalized emotional vocabulary</td>
</tr>
<tr>
<td>Wolter, Brent</td>
<td>Effects of L1 transfer on L2 learners' VN collocational use: A corpus-based study from semantic preference and semantic prosody perspective (*)</td>
</tr>
<tr>
<td>Wolvengrey, Arok</td>
<td>Expanding the coverage of a computational model for an endangered language with a derivational component – the case of Plains Cree</td>
</tr>
<tr>
<td>Woźniak, Michał</td>
<td>If an atom is a letter, then a molecule is a word: applying corpus linguistic methods to chemistry (*)</td>
</tr>
<tr>
<td>Woźniakowski, Łukasz</td>
<td>‘Que(e)rying’ identity: Intersections of straight and gay male sexual identities in ‘casual encounters’ ads (*)</td>
</tr>
<tr>
<td>Wright, David</td>
<td>Understanding street harassment of children: identifying recurrent behaviours in a corpus of young people’s accounts of harassment</td>
</tr>
<tr>
<td>Wu, Tsungming</td>
<td>Form and Function in Native and EFL Learners’ Academic Writing: What is the Difference?</td>
</tr>
<tr>
<td>Xu, Jiajin</td>
<td>Verb–Argument Constructions in Chinese EFL Learner’s Oral Production</td>
</tr>
<tr>
<td>Xu, Xiaoyu</td>
<td>An analysis of stance and voice in Applied Linguistics research articles across Mainland Chinese and British cultures</td>
</tr>
<tr>
<td>Yang, Erhong</td>
<td>A Parallel Corpus–based Study of Chinese Arabic verb phrase alignment</td>
</tr>
<tr>
<td>Yao, Xinyue</td>
<td>Exploring Grammatical Colloquialisation in Non–Native English</td>
</tr>
<tr>
<td>Zainal Abidin, Nur Najah Radhiah</td>
<td>A longitudinal learner corpus study of syntactic complexity in L2 writing</td>
</tr>
<tr>
<td>Zangerle, Eva</td>
<td>Geolocating German on Twitter – Hitches and Glitches of Building and Exploring a Twitter Corpus (*)</td>
</tr>
<tr>
<td>Zasina, Adrian Jan</td>
<td>Multi–dimensional analysis of Czech. Pilot study</td>
</tr>
<tr>
<td></td>
<td>--- Premodifying female and male adjectives in journalistic texts. A gender corpus analysis in Czech (*)</td>
</tr>
<tr>
<td>Zawisławska, Magdalena</td>
<td>Building a Polish Corpus of Synesthetic Metaphors</td>
</tr>
<tr>
<td>Zhen, Fengchao</td>
<td>Exploring the Integration of Valency Patterns and Meaning</td>
</tr>
<tr>
<td>Zinn, Jens</td>
<td>Changing Discourse–Semantics of Risk – A historical analysis of US newspapers after World War II utilizing corpus linguistics research instruments</td>
</tr>
<tr>
<td>Zottola, Angela</td>
<td>A corpus–based discourse analysis of the representation of trans people in the British press from 2013 to 2015</td>
</tr>
</tbody>
</table>
From a Plate of Spaghetti to a Cable-stayed Bridge: Increasing the Impact of Corpus Linguistics in Disciplinary Education
Susan Conrad (Portland State University, Oregon, USA)

In the 1980s, John Sinclair was instrumental in showing the profound impact corpus linguistics could have on our understanding of language. Now, ten years after his death, I want to urge corpus linguists to think again about having an impact – this time on fields that most people don’t associate with language study, such as engineering.

Why does an engineer need corpus linguistics? How can corpus-based studies improve engineering education? What does it take to move from language descriptions to applications that encourage changes in what people do? What challenges face corpus linguists in working with professionals who don’t “speak linguistics”? These are the general questions I will address, using my work in the Civil Engineering Writing Project as a concrete example.

Begun in 2009, the Civil Engineering Writing Project is a corpus-based project that addresses a long-standing problem in engineering education: students' lack of preparation for writing in the workplace. Despite decades of discussion, there had been almost no empirical investigation of the problem in the United States. I immediately saw the role corpus linguistics could play in defining the problem, informing teaching materials, and assessing improvements. The project materials have now been piloted at four universities, with significant improvements in students’ writing.

My talk will include examples of the corpus-based analyses of words and grammar that helped us understand the gaps between student and practitioner writing. The analyses have, for example, clarified the highly controversial areas of passive voice and first person pronoun use, and highlighted the importance of clausal simplicity and certain word choice issues. They demonstrate that language choices are fundamental to effective engineering. However, the linguistic analyses have also become intertwined with techniques that are less typical in corpus studies. We maintain ongoing collaborations with professionals in the community, to mine their context expertise and get their help interpreting the linguistic findings. We interview students to gain insight into reasons behind their language patterns – insights that no amount of corpus analysis can reveal. We have made additions to the research methodology to include judgments of writing effectiveness, a transition from description to evaluation that is necessary for an applied project. And we are constantly seeking new ways of turning corpus analyses into information and practice that engineers value. Although the additional techniques increase the complexity of the project, I argue in this talk that expanding corpus research in these ways can make it more useful in more disciplines.

I will reflect on the successes and the continuing challenges of the project. How exactly the plate of spaghetti and the cable-stayed bridge figure in – well, that will become clear in the talk.
Exploratory analysis of word frequencies across corpus texts: towards a critical contrast of approaches

Andrew Hardie (Lancaster University, UK)

A recent trend in corpus linguistics is the adoption of latent Dirichlet Allocation (LDA), already widely used by digital humanists (Blevins, 2010; Underwood, 2012) as a method for exploratory corpus analysis. LDA is a machine-learning approach to inducing structure in the content of a corpus based solely on word occurrence across texts or documents as data objects, one of a range of approaches usually if potentially misleadingly dubbed topic modelling. However, adopting this approach to the many-dimensional data of word frequency comes with a high price tag in terms of knowledge that the system ignores or makes nontransparent. The question this raises is whether that price tag is justified.

Various advantages have been asserted for LDA, albeit not without caveats (see Blei, 2012 for a selection of both). All such advantages notwithstanding, LDA has at least three substantive disadvantages. First, it is nondeterministic: randomisation is central to the algorithm. This is problematic from the perspective of scientific replicability for reasons too obvious to belabour. Second, its operation is opaque: the relationship between the underlying distribution data and the resulting statistical model is nontransparent to the analyst. Third, the theory of text generation underpinning the LDA algorithm is dubiously compatible with linguistic understandings of text, topic and discourse.

Moreover, although the lack of linguistic knowledge used in the construction of the model is presented as an advantage of LDA, this is equally characterisable as a disadvantage: the field of corpus analysis has invested much effort in the creation of precisely the knowledge resources which LDA is lauded for not requiring. What exactly does our acceptance of these disadvantages buy us? In examining this issue, we must venture comparisons to longer-established exploratory multivariate analysis approaches that are longer-established in corpus linguistics (cf. Biber, 1988, 1989).

Using example data drawn from the FLOB corpus, I will compare and contrast outcomes of different analytic procedures including LDA models and alternative approaches, with two questions in mind. First, to what extent are these outcomes compatible with one another? Second, to what extent are they transparently interpretable in linguistically meaningful terms?

References

Downsizing and upgrading: Why we need more spoken, more multilingual and more nonstandard corpora

Christian Mair (University of Freiburg, Germany)

Today, students of English (and a few other mostly European languages) are privileged in that they can rely on extremely rich corpus-linguistic working environments. In a brief review of 50 years’ corpus-linguistic research I will demonstrate how the availability of increasingly large corpora and increasingly sophisticated tools for analysis has left a profound mark on the discipline of linguistics. Traditional descriptive work can now be carried out to higher empirical standards. More importantly, new areas of linguistic inquiry have been opened up to rigorous empirical investigation, and corpus-based research has given a general boost to usage-based theoretical frameworks of all kinds.

As I will show, however, the story of the past fifty years has not been one of undiluted progress and success. It seems that a “conspiracy” of technological and ideological factors has favoured the creation of large monolingual standard written corpora. Data which does not fit this template tends to be made to conform to it. For example, much corpus-based work on spoken English is based on transcriptions rather than the original audio or audiovisual recordings. Similarly, complex multilingual realities tend to be simplified in corpus-compilation, for example by annotating code-switches into other languages as “extra-corpus material.”

Today, corpus technology and corpus-linguistic theorising have advanced to such an extent that these biases can and should be redressed. In the digital textual universe in which the humanities and social sciences are all operating today, the classic definition of the corpus, as a usually digital database compiled by linguists for the purposes of linguistic analysis, has become increasingly difficult to uphold and corpus-linguistics will sooner or later merge with the digital humanities movement. A kind of corpus-linguistics which emphasises spoken, multilingual and nonstandard data more than has been the case in the past will make a richer contribution to this development.
Just what is corpus stylistics?
Dan McIntyre (University of Huddersfield, UK)

Over a relatively short period of time, corpus linguistic methods have been embraced by a wide range of sub-disciplines of linguistics (and, more recently, by other disciplines entirely). Corpus linguistics has had a transformative effect on such areas as historical linguistics, child language acquisition and critical discourse analysis, to name but a few. In stylistics, corpus methods are increasingly being adopted, not least because of the influential work of corpus linguists such as Stubbs (2005) and Mahlberg (2013). Indeed, such is the popularity of the corpus approach in stylistics that it is now common to see the term corpus stylistics used to describe any stylistic work that utilises corpus methods. This adoption of corpus as a premodifier to designate a particular type of stylistics is unusual when compared against the practices of other sub-disciplines that use corpus methods. So just what is corpus stylistics and how, if at all, does it differ from corpus linguistics? My talk aims to offer answers to these questions by exploring how stylisticians have used corpora in their work. I begin with an overview of research in corpus stylistics before going on to consider issues with the presuppositions inherent in some definitions of the term. I then discuss topics in stylistics that have benefitted particularly from corpus methods. These include the analysis of speech and thought presentation (e.g. Semino et al., 1997, Semino & Short, 2004), where corpora have enabled the discovery of quantitative as well as semantic norms. Following this, I consider the washback effects that corpus linguistics has had on methodological practices in stylistics. I illustrate some of these by introducing a software tool called Worldbuilder, developed by linguists and computer scientists at the University of Huddersfield to provide a means of improving the systematicity of cognitive stylistic analyses that utilise Text World Theory (Werth, 1999). I suggest that the incorporation of basic principles from corpus linguistics such as data sampling and annotation are improving methodological and analytical practice in stylistics. Finally, having outlined the impact of corpus linguistics on stylistics, I consider what stylistics has to offer to corpus linguistics. I suggest that foregrounding theory, arguably the cornerstone of stylistics, offers valuable analytical insight when connected to notions of statistical salience.

References

Many of us are using LexisNexis, Factiva or other online sources, often in order to study a specific topic within such overall fields as gender studies, journalism, history, sociology, medicine, psychology, law.

Among the issues raised by such downloads as supplied by an online search-engine, there are choice of search-terms, duplicate articles, repeated sections within articles, online comments and discussion, disparities in formatting. But the main aim of the presentation is to focus on the problem of relevance: many of the articles retrieved may have a merely incidental mention of the desired topic.

The main aboutness of such articles doesn’t really include the topic but concerns another, quite different one. For example an article returned by a search on Brexit (Guardian, 12 January 2017) which concentrates on problems in the UK’s the National Health Service, contrasting these problems incidentally with the “theoretical risks of Brexit” and claims deficiencies in the Health Service are very obvious to ordinary voters. Its aboutness does includes Brexit but at a very minor level.

The question we will be considering is then, how do we filter aboutness so as to reduce unwanted dross? There are various aspects of relevance to identify in order to find ways of filtering out irrelevance. One concerns identifying carefully what we are really seeking in the first place, since almost any topic such as climate change, austerity, Brexit has numerous aspects (legal, social, geographical etc.), some of which are more central (within the field of knowledge) than others (gardening, hill-walking, DIY). Once it is clear which aspect of our topic is wanted, means have to be found to get rid of the others. Easier said than done!
Can Lexical Priming be detected in conversation turn-taking strategies?
Michael Pace-Sigge (University of Eastern Finland, Finland)

Turn-taking strategies in spoken communication have been widely researched and discussed in recent literature (see, e.g. McCarthy, 1998, Myers: 2009; Archer et al, 2012). Taking turns refers to when the first speaker indicates that they have finished their part of the conversation and subsequently open the floor for another speaker and, conversely, for a conversation partner to realize they can start speaking without interrupting or appearing to be impolite.

There are many prosodic, non-lexical indicators used, such as the the phonetic pointers employed by English speakers (cf. Yngve, 1970; Duncan, Jr. 1972; Knowles, 1987; Rod Gardner, 2001). Furthermore, the issue has been addressed with reference to pragmatics (see, amongst others, McCarthy, 1998; Myers: 2009; Archer et al, 2012). While anecdotal evidence, careful listening in fact, might hint at what constitute markers within conversation, a more empirical approach can be found where corpus-based evidence is used as a point of reference, as in Tao (2003), Carter & McCarthy (2006), and O’Keeffe et al (2007), and others.

This paper will look in how far the concept of textual colligation (cf. Hoey, 2015, p.13) can be applied to the production of casually spoken material. It must be noted that textual colligation has so far only been demonstrated for written texts, though it is predicted to apply to turn-taking (see McCarthy, 2010 or Evison, 2012, for details). Hoey links the subconscious process of textual colligation to the phenomenon known as Lexical Priming – where speakers and listeners are primed through repeat exposure (cf. Hoey, 2005; Pace-Sigge, 2013) of a word or set of words: primes signal what can be expected next.

When looking at evidence of signals of turn-endings or turn-starts, however, the nature of spoken language presents a number of clear difficulties (cf. Halliday, 2004). The issue becomes difficult when one tries to find clear lexical signals of either turn-givers or turn-takers in a transcript. While there are a large number of turns between speakers in any conversation, there is no direct correlation between a clearly defined set of words and the length of turns. This is partly due to the mental processing power of any speaker and the online nature of conversations (cf. Cheng, 2012, pp. 13f.) and partly due to the nature of free-flowing, casual dialogue between (fairly) equal parties, which follows no fixed norm (unlike formal dialogues or conversations between unequal parties – for example doctor-patient, teacher-student).

Yet, following the tenets of psychological priming that form the basis of Hoey’s lexical priming theory, some kind of trigger item should be in evidence, showing a listener that a turn is given up. Levinson and Torreira suggest a form of “long range prediction in comprehension” (2015). Likewise, the first speaker should receive a notification whether the listener is prepared to either forgo or take the turn. When we look at the lexis employed, informal speech appears, on first sight, too anarchic to employ a formulaic, standardised protocol. In order to facilitate smooth and fluent conversation, turn-taking must, however, follow a structured recognisable pattern.

Following research presented by Evinson (2012), this paper will a) describe some salient signals used that become apparent when monologues are directly compared with dialogues. The paper then focuses on b) lexical items that appear to be prevalent triggers
to hand over a turn or markers of turn-starts. The research presented here shows preference or dispreference for select categories of words and key items. Key items have been found to be prevalent at the start and end of turns; these items appear with a statistically significantly higher rate of occurrence.

Based on this it will be argued that turn-taking does appear to follow a structured, recognisable pattern, lexically, in order to facilitate conversation.

The material used for this investigation will be exclusively British English spoken data. The corpora used for this investigation are the section of the SCO Corpus (2013) which has clearly identifiable turns, the Lancaster SWAT corpus (2003) and the 2009 “Linguistic Innovators Corpus” (LIC) of young and older speakers in both Hackney and Havering (i.e. North London and a part of ‘Greater London’). This third sub-corpus presents the largest part of the data and the largest number of turns identified.

References

Evison, J. (2012). A corpus linguistic analysis of turn-openings in spoken academic discourse: Understanding discursive specialisation. English Profile Journal, 3(4). Link to this article: http://journals.cambridge.org/abstract_S2041536212000049 (last accessed 05.10.15)
O’Keeffe, Anne, Michael McCarthy and Ronald Carter (2007). *From Corpus to Classroom. Language use and language teaching*. Cambridge: CUP

**Corpora**


Material kindly provided by Paul Kerswill.
Introduction

The study is concerned with investigating religion in all British broadsheet newspapers in two periods of time (January 2001 and January 2010). The purpose of choosing such periods is because of the global awareness of religion that followed the events occurred in the first decade of 21st century. Recent studies (Weller, 2011) have shown the place and the impact of religion on the public in Britain in the first decade of this century because of a number of events such as the New York 9/11 attacks of 2001, the Madrid bombing 11/3 of 2004, the London 7/7 bombing of 2005, and the Glasgow 30/6 bombing of 2007, as well as the invasions of Afghanistan in 2001 and Iraq in 2003 and other events. Thus religion has radically changed and has broken into mainstream Western news agendas (Gower and Mitchell, 2012).

I intend to analyse the press representation of three widespread religions in the UK – Christianity, Judaism, and Islam as it is impossible to cover all world religions in one research study. These three monotheistic religions are mostly practiced in the UK according to the census data in 2011 in the UK (ONS 2011); Christianity occupied the majority; Islam is next, and then Judaism.

Investigating religions has been the focus of a range of frameworks in different disciplines such as media, theology, and sociology, but it has not been tackled with within the area of linguistics. Accordingly, this study is conducted using a selected set of corpus linguistic tools namely frequency lists, concordances, and collocations.

I present a case study investigating applicability of the corpus tool Sketch Engine to sort collocates of those three religions. I aim to examine language used to represent these religions and their followers as well, i.e. how many times a certain group have been represented in the press as grammatical agents or patients in order to illustrate any changes in representation of religion over time.

Data collection


The data consists of 1032 and 1168 articles in 2001 and 2010 respectively after excluding all duplicated and unrelated articles. All articles were then uploaded into Sketch Engine to conduct a corpus analysis. The number of words in each corpus is 797,336 in 2001 and 994,081 in 2010.
As far as these three Abrahamic religions are concerned, I intend to view the raw frequency (RF), which needs to be normalised based on the normalised frequency (NF) per 1.000 words as the size of the two corpora is different. McEnery et al. (2006: 52-3) state that when comparing two corpora of different sizes, raw frequencies extracted from corpora need to be normalised to a common base. The following table show the raw and normalised frequencies for each search query:

<table>
<thead>
<tr>
<th>Words</th>
<th>2001/RF</th>
<th>NF</th>
<th>2010/RF</th>
<th>NF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Christianity</td>
<td>79</td>
<td>0.09</td>
<td>92</td>
<td>0.09</td>
</tr>
<tr>
<td>Christians</td>
<td>87</td>
<td>0.10</td>
<td>162</td>
<td>0.16</td>
</tr>
<tr>
<td>Christian</td>
<td>537</td>
<td>0.67</td>
<td>644</td>
<td>0.64</td>
</tr>
<tr>
<td>Judaism</td>
<td>48</td>
<td>0.06</td>
<td>25</td>
<td>0.02</td>
</tr>
<tr>
<td>Jews</td>
<td>406</td>
<td>0.50</td>
<td>240</td>
<td>0.20</td>
</tr>
<tr>
<td>Jew</td>
<td>253</td>
<td>0.25</td>
<td>149</td>
<td>0.14</td>
</tr>
<tr>
<td>Jewish</td>
<td>677</td>
<td>0.84</td>
<td>446</td>
<td>0.44</td>
</tr>
<tr>
<td>Islam</td>
<td>64</td>
<td>0.08</td>
<td>229</td>
<td>0.23</td>
</tr>
<tr>
<td>Muslim</td>
<td>295</td>
<td>0.36</td>
<td>883</td>
<td>0.88</td>
</tr>
<tr>
<td>Muslims</td>
<td>123</td>
<td>0.15</td>
<td>380</td>
<td>0.38</td>
</tr>
<tr>
<td>Islamic</td>
<td>231</td>
<td>0.28</td>
<td>453</td>
<td>0.45</td>
</tr>
</tbody>
</table>

**Table (1):** The raw and normalised frequencies of the search items

From the table, it is apparent that the frequency of the search items is varied in both periods of time in which some decreased, while others increased in 2010. For instance, ‘Judaism’ is the least search word in both periods of time; ‘Christianity’ kept the same level of frequency, while ‘Islam’ increased considerably (0.08 to 0.23). It is worth mentioning that the singular forms of some search words in both corpora tend to occur as a noun and an adjective as in the case of Christian and Muslim. As a result, the singular form of Christian and Muslim in both periods is higher than its plural form as shown in the table above, but the adjectival pattern Jewish, is declined sharply in 2010 (0.84 to 0.44). words 2001/RF NF 2010/RF NF Christianity 79 0.09 92 0.09 Christians 87 0.10 162 0.16 Christian 537 0.67 644 0.64 Judaism 48 0.06 25 0.02 Jews 406 0.50 240 0.20 Jew 253 0.25 149 0.14 Jewish 677 0.84 446 0.44 Islam 64 0.08 229 0.23 Muslim 295 0.36 883 0.88 Muslims 123 0.15 380 0.38 Islamic 231 0.28 453 0.45

Regarding the followers or believers of these three religions, in the first place the singular form of Christian increased in 2010 in terms of raw frequency 537 to 644, but decreased slightly in terms of normalised frequency (0.67) to (0.64) though the size of the corpus in 2010 is larger than that in 2001. However, the plural from Christians rose from 87 (0.10) to 162 (0.16). In addition, Muslim trebled in 2010 (0.36 to 0.88), and its plural form Muslims doubled (0.15 to 0.38). Jew or Jews decreased in 2010 in terms of both raw and normalised frequencies.

**Word sketches**

Having outlined the frequency differences of the items under investigation, I intend to report the findings of word sketches that I carried out for each search word. I uploaded my data into Sketch Engine – a corpus query tool developed by Kilgarriff et
al. (2004) to identify collocates in a range of grammatical relations via Word Sketch. Sketch Engine is an important and useful tool to analyse the grammatical features of salient collocates. It not only identifies collocates, but also specifies the grammatical constructions within such collocates (Baker et al. 2013: 37). Therefore I used Sketch Engine to create a word sketch for each search query. Word Sketch is a one-page automatic, corpus-based summary of a word’s grammatical and collocational behaviour, (Kilgariff et al. 2004). It provides a summary of the main part of speech of collocates of the target word.

Creating a word sketch for each search word, i.e. Christianity, Christian, Judaism, Jew, Jews, Jewish, Islam, Muslim, Muslims, and Islamic, I investigated the grammatical relations between collocates which can tell me something about phenomena like agency. Word Sketch provides a set of significant collocates of each item sorted into various grammatical frames or patterns. For instance, verbs and adjectives co-occur with the search query might reveal how a certain group such as Christians, Jews, and Muslims, are presented as agents or goals, and how they are described in the British newspapers. This therefore leads me to shed some more light on the nature of agency extracted from the sorts of verbs collocating with the search items. Further Christian, Jewish, Muslim, and Islamic, as adjectives, can identify the nouns that are collocating with them, and so reveal which sorts of nouns such items describe. Therefore, a special attention is paid to three grammatical patterns: subjects, objects, and modifiers.

The following table shows the number of occurrences of each item in the word sketch plus its statistics in both periods. The statistic score is measured by logDice based on how salient the words are. Salience is the basic measure of Word Sketch. Moreover, LogDice is based on Dice coefficient which reveals the more frequent lexical collocates (For more information, see Rychlý, 2008).

<table>
<thead>
<tr>
<th>Items</th>
<th>On. of occurrences/2001</th>
<th>Statistics pm</th>
<th>On. Of occurrences/2010</th>
<th>Statistics pm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Christianity</td>
<td>79</td>
<td>82.48</td>
<td>92</td>
<td>76.95</td>
</tr>
<tr>
<td>Christian (n.)</td>
<td>282</td>
<td>294.43</td>
<td>311</td>
<td>260.15</td>
</tr>
<tr>
<td>Christian (adj.)</td>
<td>255</td>
<td>266.24</td>
<td>333</td>
<td>278.55</td>
</tr>
<tr>
<td>Judaism</td>
<td>48</td>
<td>50.11</td>
<td>25</td>
<td>20.19</td>
</tr>
<tr>
<td>Jew</td>
<td>250</td>
<td>261.02</td>
<td>149</td>
<td>124.64</td>
</tr>
<tr>
<td>Jews</td>
<td>233</td>
<td>243.27</td>
<td>140</td>
<td>117.11</td>
</tr>
<tr>
<td>Jewish</td>
<td>635</td>
<td>663.00</td>
<td>421</td>
<td>352.17</td>
</tr>
<tr>
<td>Islam</td>
<td>64</td>
<td>66.82</td>
<td>229</td>
<td>191.56</td>
</tr>
<tr>
<td>Muslim (n.)</td>
<td>119</td>
<td>124.24</td>
<td>344</td>
<td>287.76</td>
</tr>
<tr>
<td>Muslim (adj.)</td>
<td>179</td>
<td>183.76</td>
<td>539</td>
<td>450.88</td>
</tr>
<tr>
<td>Muslims</td>
<td>72</td>
<td>75.17</td>
<td>197</td>
<td>164.79</td>
</tr>
<tr>
<td>Islamic</td>
<td>179</td>
<td>186.89</td>
<td>325</td>
<td>272.16</td>
</tr>
</tbody>
</table>

Table (2): Number of occurrences of the search items in the word sketch in 2001 and 2010

In terms of statistics, there is a decline in the findings regarding Judaism, Jew (s), and Jewish, while it is the opposite in the case of Islam, Muslim (s), and Islamic in 2010. Christianity and Christian kept almost the same statistic scores in both periods.
By way of example, Jews decreased from (243.27) in 2001 to (117.11) in 2010, Muslims increased from (75.17) in 2001 to (164.79) in 2010. Further a word sketch viewed that the result of Christian as a noun and an adjective is somewhat distinct (294.43) and (266.24) respectively. But Jewish is more frequent and salient than the noun Jew (s) (table 2). Similarly, Muslim tended to be used as an adjective (183.76) more frequently than as a noun (124.24).

It is worth mentioning that word sketches of Christian, Jew, and Muslim tagged as singular nouns display also their plural forms; thus I consider the plural forms alone to find out how such groups (Jews and Muslims) are represented and behave. However, the sketch has not shown any significant collocates in the grammatical features in the case of Christians in both periods; only 3 (3.13 per million) and 6 (5.01 per million) in 2001 and 2010 respectively. As a result, I examined Christians in connection with the singular form as the word sketch encompasses both forms of the target word.

References


Lexis of Spanish video game magazines: preferred mechanisms to adapt anglicisms
Dr Carmen F. Cabrera Álvarez (ULPGC, Spain)

The world of video games in Spain can be highly influenced by the use of anglicisms. The language and style used in some Spanish video game magazines were analysed in this paper. A list of anglicisms was extracted and classified into eight different categories (Brito Pérez, 2002) in order to discover which are the preferred mechanisms to adapt these words into Spanish.

A corpus of 1,195 anglicisms was found in the Spanish video game magazines that were analysed (Micromanía, Playmanía and Revista Oficial Nintendo). This confirms there is a significant amount of anglicisms in this type of specialised press. Calques were the most recurrent mechanisms: 22% of these words were creation calques (some words in Spanish are combined to create a new meaning: high definition, alta definición) and 14% were extension calques (a new meaning is given to a word that already exists in Spanish: the cloud, la nube). There was also a significant amount (23%) of hybrid compound anglicisms (these words combine Spanish and English words: medios online) and pure anglicisms (18%) that have not been altered at all and are used in their original form in Spanish: survival horror.

The use of anglicisms in video games is not only due to the influence of the English language in the technology field (Brito Pérez, 2002), but also to the fact that so many concepts are created so fast that it is difficult to think of a way to adapt those words into Spanish. There is also the social factor, as English words look more technical (Brito Pérez, 2002) and they even sound more modern and young (Gómez Capuz, 1998) in this type of articles.

After analysing the data, I found out that the use of anglicisms was not always consistent (Cabrera Álvarez, 2015). More than one type of anglicisms could be used for the same concept. However, one of these options was always used more often. For example, seguidor was used 5 times in all the magazines that were analysed, whereas fan was used 39 times.

I observed 3 different options to decide which term to choose: the word that was easier to write or to pronounce was used (juego de acción y rol instead of Action RPG); the word that has been used in Spanish for a very long time was preferred, even if it has not been completely adapted (fan and seguidor); and any of the options was used as the concept is so new, it is not known how it is going to evolve (cool hunting and captura de tendencias) (Cabrera Álvarez, 2015).

Finally, there is a brief description on how the use of anglicisms influences video game titles. I discovered that 84% of 624 video games had their title in English with no modifications (Cabrera Álvarez, 2015). Some of these video games were developed in non-English speaking countries.

A glossary with video game terms can be created in the future to help experts in the field and translators normalise the use of these words.
References


The Negation Operator is not a Suppressor of the Concept in its Scope. In Fact, Quite the Opposite
Israel Becker (Tel-Aviv University, Israel)

Introduction

The effect of the negation operator (henceforth, negator) on the activation levels in memory of the concept in its scope is a controversial topic: Some argue that the negator unconditionally reduces the initial activation levels of the concept in its scope to baseline levels or below, thereby assigning to the negator the role of a suppressor (e.g., Kaup, Lüdtke, & Zwaan 2006; MacDonald & Just 1989). Others argue that the initial activation levels of the concept are not automatically suppressed by its negator. As a result, the negated concept is retained in memory (e.g., Giora 2003; Giora, Balaban, Fein, & Alkabets 2005; Giora, Fein, Aschkenazi, & Alkabets-Zlozover 2007).

The current study aims to provide a corpus-based support to the functional Retention Hypothesis by looking at a corpus of natural speech, thus rejecting the unconditional Suppression Hypothesis.

Tracking 'suppression' or 'retention' of negated concepts in a corpus

'Suppression' and 'retention' of concepts (indicated by their activation levels in memory) are online processes, and as such are monitored via online experiments, that is, experiments in which the reaction time of participants to tasks related to a negated concept is measured. Corpora, unlike online experiments, are offline entities (i.e., they involve no real-time measurements). As such, corpora can only point at traces of 'suppression' or 'retention' of negated concepts, thus providing only indirect support to any of the aforementioned hypotheses. How, then, can one track traces of 'suppression' or 'retention' of negated concepts in a corpus?

Given that the processing of concepts affects (at least, to some extent) their interpretation, results of offline and online experiments using the same materials can indicate a possible link between the way a negated expression is represented and the activation status of a negated concept.

A lead as for how the representation of a negated expression attests to the activation levels of a negated concept is provided by Giora et al. (2005, 2007) who conducted online and offline experiments using the same materials. In the online experiments, they showed that the initial activation levels of negated concepts (e.g. happy in not happy) are not any different from the activation levels of their affirmative counterparts (e.g., happy). Results of the offline experiments showed that raters opted for a mitigated version of the antonym of the negated item: They considered not happy as a mitigated version of sad. Taken together, these results suggest that the negator is a weak operator and as such it retains the concept in its scope activated in memory, which is, possibly, the reason why the negated expression is rated as a mitigated version of the antonym of the negated item and it does not assume the uttermost end of a conceptual-argumentative scale (e.g., not sad≠happy).
Accordingly, one should look into corpora for instances of conceptually-argumentatively weak negative expressions and examine the ways in which the corpus echoes the negated concepts – whether the negated concept is retained or suppressed.

**The specific discourse pattern to be looking for in a corpus in order to test the Hypotheses**

Instances of multiple negated utterances — Resumptive Negation — seem like good candidates to test the abovementioned hypotheses, as they seem to be the consequence of the "weak" nature of negators, as argued by Jespersen (1917:72-75). But in order to narrow the search for Resumptive Negation patterns which indicate that the presumably "weak" character of the negator is what requires additional negators, one should apply Du Bois' (2014) Principles of Dialogic Syntax according to which a selective addition of a morphologically-distinct negator has to do with some "trouble" with the original negator.

Such a discourse pattern, which is the focus of the current study, is exemplified in (1) extracted from a 1990s episode of a nightly talk-show, where Larry King, the host, and his guest, Donald Trump, discuss Trump's upcoming divorce. Trump denies King's insinuation about his inclination to play the role of the ultimate playboy rather than having a monogamous relationship:

(1) [...] so I'm not one that loves the concept of divorce.
   In fact, just the opposite,
   I hate the concept of divorce,
   I hate everything it represents.
   There is nothing better than a good marriage.

(27.7.1990)

Note that the unmarked negator not in (1) is followed by the morphologically-distinct marked negator the opposite.

I name this discourse pattern the Resumptively-Negated Denial Pattern (henceforth, RNDP). Figure 1 below is a schematic form of the RNDP, which is exemplified by using the specifics of (1).

<table>
<thead>
<tr>
<th>(i)</th>
<th>A concept to be denied in (a)</th>
<th>love the concept of divorce</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>A weakly negated version of the previous concept</td>
<td>I'm not one that loves the concept of divorce</td>
</tr>
<tr>
<td>(b)</td>
<td>A connective that implies that the speaker commits that (c) is a stronger claim than (a)</td>
<td>In fact</td>
</tr>
<tr>
<td>(c)</td>
<td>A stronger version of (a)</td>
<td>Just the opposite</td>
</tr>
<tr>
<td>(ii)</td>
<td>An affirmative spell-out version of (c)</td>
<td>I hate the concept of divorce</td>
</tr>
</tbody>
</table>

**Figure 1.** Schematic form of the RNDP exemplified by using the specifics of (1)
The dataset

Since the RNDP is a long sequence and as such it is not expected to occur frequently in corpora, only very large corpora should be considered. These corpora should necessarily be spoken corpora as samples of spoken discourse contain as twice as many instances of negation as samples of written discourse (of the same size) (Tottie 1991:17). As attested from example (1) above, the RNDP is an outcome of interlocutors' intensive mutual monitoring during discourse which is, in turn, the outcome of a direct social interaction between interlocutors (Goffman 1964). Therefore, large spoken corpora documenting direct social interactions between co-participants should be focused on. Such corpora are expected to contain (relatively) many instances of the RNDP.

My dataset was, therefore, extracted out of the spoken section of COCA (Davies 2008), which contains ~95M tokens from transcripts of face-to-face and telephone conversations recorded from (American) TV and radio programs between 1990-2012. I first extracted all instances of the marked negators, the opposite and the contrary, from which I further extracted all instances of the RNDP in which the opposite or the contrary follow an unmarked negator such as no, not, never, none, -n't, etc. (Tottie 1991:8), altogether 399 instances of the RNDP.

The negative expression in Part (a) is a mitigated version of Part (c)

Recall that the first step in the current analysis is to establish that Part (a) of the RNDP is a weak negative expression which does not assume the uttermost end of a conceptual-argumentative scale. Such a Part (a) is expected to be the outcome of a weak negator, such that does not suppress the concept in its scope.

Part (b) of the RNDP which accommodates quite often an emphasizing connective (e.g., in fact, actually, or indeed), shows that Part (a) is, indeed, construed by the speaker as a mitigated negative expression. This emphasizing connective indicates that the succeeding utterance (i.e., Part (c)) is stronger than the preceding utterance (i.e., Part (a)) both conceptually (e.g., Horn 1989:231-252) and argumentatively (e.g., Schwenter & Traugott 2000), and therefore cannot assume the uttermost end of a conceptual-argumentative scale.

The concept in the scope of the negator (in Part (a)) remains accessible in memory

If Part (a) does not assume the uttermost end of a conceptual-argumentative scale, then according to the prediction detailed earlier, the RNDP should display a highly accessible negated concept. And this is indeed the case: all 399 instances (but 5) of the RNDP display a zero anaphora in the scope of the additional negator, the opposite/contrary $t\ddagger$, co-referencing to the concept in the scope of the preceding negator. According to the universal cognitive Accessibility Theory (Ariel 1990 and onwards), a zero anaphora indicates the highest possible mental accessibility of a concept, thus attesting to a high mental accessibility of the negated concept rather than its unconditional suppression.
Further analysis of the RNDP as a repair

The RNDP is further analyzed along the lines suggested by van der Wouden (1997), who contended that multiple negation constructions are instances of an apposition construction and as such, their "first aim [...] is self-correction" (van der Wouden 2000:240). This analysis reveals that the RNDP is a self-repair. Further analysis shows that as a self-repair, the RNDP is an appropriateness-repair rather than an error-repair (Levitt 1989). This corpus evidence taken together with the findings of Shuval & Hemforth (2008), who conducted eye-fixation experiments and found that negated concepts in appropriateness-repair constructions were significantly more accessible than negated concepts in error-repair constructions, show that the negator in Part (a) of the RNDP cannot possibly suppress the concept in its scope.

Summary

In this corpus-based study I present complementary evidence from natural speech to psycholinguistic results, showing that the concept in the scope of a negator is retained in memory and is not suppressed unconditionally. I suggest that a conceptually-argumentatively weak negative expression is the outcome of an activated concept in the scope of a negator. Having traced a discourse pattern accommodating a negative expression which does not assume the outermost end of a conceptual-argumentative scale, I show that this discourse pattern consistently manifests a highly accessible concept in the scope of the negator. I then present an independent line of analysis which combines corpus data, attesting to the appropriateness-repair character of the RNDP, with psycholinguistic findings of the processing of negated concepts in an appropriateness-repair, showing that as an appropriateness-repair the RNDP must retain the negated concept active in memory.

References


The Asymmetry of Conceptual Expansions in Predicational and Modificational Contexts
Sadayuki Okada (Osaka University, Japan)

Body-part terms are known to exhibit a wealth of extended (metonymical and metaphorical) meanings. Here I take up 16 English body-part terms and collect 250 argument usages and the same number of adjunct usages from the BNC for each nominal expression, and sort out the extended references to demonstrate an asymmetrical distribution: that is, part of the extended references found in argument positions are attested in adjuncts, but not the other way round (cf. Waltereit 1999). This asymmetrical distribution found in the predicational context of argument-adjunct combinations is also reflected in the modificational context of head-modifier combinations. The extended references attested in modifier positions are basically found in head positions of word combinations, but not the other way round (cf. Arcodia 2011).

These two distributions of nominal conceptual expansions are related in a number of ways, and some observations can be made regarding the interrelation. They include:

(1) Those extended meanings attested both in arguments and adjuncts freely are most likely to be attested also in both head and modifier positions in word combinations.

(2) Those meanings attested in adjunct positions are more likely to be assisted by premodification, while in argument positions they are less likely to be hosted by modifiers.

(3) Those meanings assisted by premodifiers both in argument and adjunct positions are unlikely to be employed in the position of modifiers in word combinations.

(1) shows that expanded references attested in different syntactic contexts are more entrenched and are likely to be regarded as part of the lexical meaning of the body-part terms. Instances include nose for "tip/front,", head for "leader," mouth for "aperture," face for "front/surface" etc. Those meanings are readily adopted in word combinations, and even appear in modifier positions where no assistance of additional linguistic material for insinuating a particular extended meaning is expected.

(2) indicates that nominals in argument positions are not likely to require additional modifiers for decoding the extended meaning, because selectional restrictions dictated by the main predicate of the sentence is enough to identify the intended sense, while in adjuncts, additional clues tend to appear for deciphering. e.g., Her browdarkened. (argument use of brow for "complexion") vs. He looks forward to the future, with a stem brow (adjunct use of brow for "complexion")

(3) means that those extended meanings requiring modifiers for reaching the appropriate explication are not likely to obtain when the nominals stand alone. Heads are given semantic clues by modifiers for deciphering, but modifiers are simply restrictors in word combinations, so that the extended meanings of modifiers should obtain even when they stand alone. This can be the reason for the distribution in (3). e.g., a covetous eye (eye for "disposition expressed in the eye," which is not designated when the nominal eye stands alone. No example is attested where eye is employed as the modifier in a word combination with the meaning of "disposition.")

Some additional distributional facts will be reported through the scrutiny of BNC data, and the reasons for the distribution will be investigated.

References
Representing Religion in the British Press: A Corpus-based Analysis
Sawsan Hassan (University of Huddersfield, UK)

Introduction

The study is concerned with investigating religion in all British broadsheet newspapers in two periods of time (January 2001 and January 2010). The purpose of choosing such periods is because of the global awareness of religion that followed the events occurred in the first decade of 21st century. Recent studies (Weller, 2011) have shown the place and the impact of religion on the public in Britain in the first decade of this century because of a number of events such as the New York 9/11 attacks of 2001, the Madrid bombing 11/3 of 2004, the London 7/7 bombing of 2005, and the Glasgow 30/6 bombing of 2007, as well as the invasions of Afghanistan in 2001 and Iraq in 2003 and other events. Thus religion has radically changed and has broken into mainstream Western news agendas (Gower and Mitchell, 2012).

I intend to analyse the press representation of three widespread religions in the UK – Christianity, Judaism, and Islam as it is impossible to cover all world religions in one research study. These three monotheistic religions are mostly practiced in the UK according to the census data in 2011 in the UK (ONS 2011); Christianity occupied the majority; Islam is next, and then Judaism.

Investigating religions has been the focus of a range of frameworks in different disciplines such as media, theology, and sociology, but it has not been tackled with within the area of linguistics. Accordingly, this study is conducted using a selected set of corpus linguistic tools namely frequency lists, concordances, and collocations.

I present a case study investigating applicability of the corpus tool Sketch Engine to sort collocates of those three religions. I aim to examine language used to represent these religions and their followers as well, i.e. how many times a certain group have been represented in the press as grammatical agents or patients in order to illustrate any changes in representation of religion over time.

Data collection


The data consists of 1032 and 1168 articles in 2001 and 2010 respectively after excluding all duplicated and unrelated articles. All articles were then uploaded into Sketch Engine to conduct a corpus analysis. The number of words in each corpus is 797,336 in 2001 and 994,081 in 2010.
As far as these three Abrahamic religions are concerned, I intend to view the raw frequency (RF), which needs to be normalised based on the normalised frequency (NF) per 1,000 words as the size of the two corpora is different. McEnery et al. (2006: 52-3) state that when comparing two corpora of different sizes, raw frequencies extracted from corpora need to be normalised to a common base. The following table show the raw and normalised frequencies for each search query:

<table>
<thead>
<tr>
<th>Words</th>
<th>2001/RF</th>
<th>NF</th>
<th>2010/RF</th>
<th>NF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Christianity</td>
<td>79</td>
<td>0.09</td>
<td>92</td>
<td>0.09</td>
</tr>
<tr>
<td>Christians</td>
<td>87</td>
<td>0.10</td>
<td>162</td>
<td>0.16</td>
</tr>
<tr>
<td>Christian</td>
<td>537</td>
<td>0.67</td>
<td>644</td>
<td>0.64</td>
</tr>
<tr>
<td>Judaism</td>
<td>48</td>
<td>0.06</td>
<td>25</td>
<td>0.02</td>
</tr>
<tr>
<td>Jews</td>
<td>406</td>
<td>0.50</td>
<td>240</td>
<td>0.20</td>
</tr>
<tr>
<td>Jew</td>
<td>253</td>
<td>0.25</td>
<td>149</td>
<td>0.14</td>
</tr>
<tr>
<td>Jewish</td>
<td>677</td>
<td>0.84</td>
<td>446</td>
<td>0.44</td>
</tr>
<tr>
<td>Islam</td>
<td>64</td>
<td>0.08</td>
<td>229</td>
<td>0.23</td>
</tr>
<tr>
<td>Muslim</td>
<td>295</td>
<td>0.36</td>
<td>883</td>
<td>0.88</td>
</tr>
<tr>
<td>Muslims</td>
<td>123</td>
<td>0.15</td>
<td>380</td>
<td>0.38</td>
</tr>
<tr>
<td>Islamic</td>
<td>231</td>
<td>0.28</td>
<td>453</td>
<td>0.45</td>
</tr>
</tbody>
</table>

Table (1): The raw and normalised frequencies of the search items

From the table, it is apparent that the frequency of the search items is varied in both periods of time in which some decreased, while others increased in 2010. For instance, ‘Judaism’ is the least search word in both periods of time; ‘Christianity’ kept the same level of frequency, while ‘Islam’ increased considerably (0.08 to 0.23). It is worth mentioning that the singular forms of some search words in both corpora tend to occur as a noun and an adjective as in the case of Christian and Muslim. As a result, the singular form of Christian and Muslim in both periods is higher than its plural form as shown in the table above, but the adjectival pattern Jewish, is declined sharply in 2010 (0.84 to 0.44). words 2001/RF NF 2010/RF NF Christianity 79 0.09 92 0.09 Christians 87 0.10 162 0.16 Christian 537 0.67 644 0.64 Judaism 48 0.06 25 0.02 Jews 406 0.50 240 0.20 Jew 253 0.25 149 0.14 Jewish 677 0.84 446 0.44 Islam 64 0.08 229 0.23 Muslim 295 0.36 883 0.88 Muslims 123 0.15 380 0.38 Islamic 231 0.28 453 0.45

Regarding the followers or believers of these three religions, in the first place the singular form of Christian increased in 2010 in terms of raw frequency 537 to 644, but decreased slightly in terms of normalised frequency (0.67) to (0.64) though the size of the corpus in 2010 is larger than that in 2001. However, the plural from Christians rose from 87 (0.10) to 162 (0.16). In addition, Muslim trebled in 2010 (0.36 to 0.88), and its plural form Muslims doubled (0.15 to 0.38). Jew or Jews decreased in 2010 in terms of both raw and normalised frequencies.

Word sketches

Having outlined the frequency differences of the items under investigation, I intend to report the findings of word sketches that I carried out for each search word. I uploaded my data into Sketch Engine – a corpus query tool developed by Kilgarriff et
al. (2004) to identify collocates in a range of grammatical relations via Word Sketch. Sketch Engine is an important and useful tool to analyse the grammatical features of salient collocates. It not only identifies collocates, but also specifies the grammatical constructions within such collocates (Baker et al. 2013: 37). Therefore I used Sketch Engine to create a word sketch for each search query. Word Sketch is a one-page automatic, corpus-based summary of a word’s grammatical and collocational behaviour, (Kilgariff et al. 2004). It provides a summary of the main part of speech of collocates of the target word.

Creating a word sketch for each search word, i.e. Christianity, Christian, Judaism, Jew, Jews, Jewish, Islam, Muslim, Muslims, and Islamic, I investigated the grammatical relations between collocates which can tell me something about phenomena like agency. Word Sketch provides a set of significant collocates of each item sorted into various grammatical frames or patterns. For instance, verbs and adjectives co-occur with the search query might reveal how a certain group such as Christians, Jews, and Muslims, are presented as agents or goals, and how they are described in the British newspapers. This therefore leads me to shed some more light on the nature of agency extracted from the sorts of verbs collocating with the search items. Further Christian, Jewish, Muslim, and Islamic, as adjectives, can identify the nouns that are collocating with them, and so reveal which sorts of nouns such items describe. Therefore, a special attention is paid to three grammatical patterns: subjects, objects, and modifiers.

The following table shows the number of occurrences of each item in the word sketch plus its statistics in both periods. The statistic score is measured by logDice based on how salient the words are. Salience is the basic measure of Word Sketch. Moreover, LogDice is based on Dice coefficient which reveals the more frequent lexical collocates (For more information, see Rychlý, 2008).

<table>
<thead>
<tr>
<th>Items</th>
<th>On. of occurrences/2001</th>
<th>Statistics pm</th>
<th>On. Of occurrences/2010</th>
<th>Statistics pm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Christianity</td>
<td>79</td>
<td>82.48</td>
<td>92</td>
<td>76.95</td>
</tr>
<tr>
<td>Christian (n.)</td>
<td>282</td>
<td>294.43</td>
<td>311</td>
<td>260.15</td>
</tr>
<tr>
<td>Christian (adj.)</td>
<td>255</td>
<td>266.24</td>
<td>333</td>
<td>278.55</td>
</tr>
<tr>
<td>Judaism</td>
<td>48</td>
<td>50.11</td>
<td>25</td>
<td>20.19</td>
</tr>
<tr>
<td>Jew</td>
<td>250</td>
<td>261.02</td>
<td>149</td>
<td>124.64</td>
</tr>
<tr>
<td>Jews</td>
<td>233</td>
<td>243.27</td>
<td>140</td>
<td>117.11</td>
</tr>
<tr>
<td>Jewish</td>
<td>635</td>
<td>663.00</td>
<td>421</td>
<td>352.17</td>
</tr>
<tr>
<td>Islam</td>
<td>64</td>
<td>66.82</td>
<td>229</td>
<td>191.56</td>
</tr>
<tr>
<td>Muslim (n.)</td>
<td>119</td>
<td>124.24</td>
<td>344</td>
<td>287.76</td>
</tr>
<tr>
<td>Muslim (adj.)</td>
<td>179</td>
<td>183.76</td>
<td>539</td>
<td>450.88</td>
</tr>
<tr>
<td>Muslims</td>
<td>72</td>
<td>75.17</td>
<td>197</td>
<td>164.79</td>
</tr>
<tr>
<td>Islamic</td>
<td>179</td>
<td>186.89</td>
<td>325</td>
<td>272.16</td>
</tr>
</tbody>
</table>

Table (2): Number of occurrences of the search items in the word sketch in 2001 and 2010

In terms of statistics, there is a decline in the findings regarding Judaism, Jew (s), and Jewish, while it is the opposite in the case of Islam, Muslim (s), and Islamic in 2010. Christianity and Christian kept almost the same statistic scores in both periods.
By way of example, Jews decreased from (243.27) in 2001 to (117.11) in 2010, Muslims increased from (75.17) in 2001 to (164.79) in 2010. Further a word sketch viewed that the result of Christian as a noun and an adjective is somewhat distinct (294.43) and (266.24) respectively. But Jewish is more frequent and salient than the noun Jew (s) (table 2). Similarly, Muslim tended to be used as an adjective (183.76) more frequently than as a noun (124.24).

It is worth mentioning that word sketches of Christian, Jew, and Muslim tagged as singular nouns display also their plural forms; thus I consider the plural forms alone to find out how such groups (Jews and Muslims) are represented and behave. However, the sketch has not shown any significant collocates in the grammatical features in the case of Christians in both periods; only 3 (3.13 per million) and 6 (5.01 per million) in 2001 and 2010 respectively. As a result, I examined Christians in connection with the singular form as the word sketch encompasses both forms of the target word.

References


In usage-based cognitive linguistics it is widely accepted that language operates on two complementing functions, namely, the symbolic and the communicative functions (Evans and Green, 2006). These functions, specially the communicative one, foreground the importance that specific usage-events and communicative intentions have in determining the linguistic structures we conceptualize. This characteristic can be evidenced in a number of constructions whose form-function relations cannot be fully accounted for without a thorough analysis of their corresponding pragmatic factors. Such is the case of indeterminate subjects, which we aim to examine in this paper.

From a discursive standpoint, pragmatic factors such as the speaker’s communicative intentions are believed to be the driving force for different grammatical constructions, that is, speakers can make use of different syntactic strategies to codify their intentions linguistically. In CxG (Goldberg, 1995), this view is also captured by the Principle of No Synonymy (Corollary B) according to which, two or more syntactically distinct and pragmatically synonymous constructions must not be semantically synonymous. We believe this analysis can comfortably accommodate constructions such as passives, actives with generic pronouns (e.g. you, one, someone, etc.) and the Brazilian Portuguese indeterminate subject construction (e.g. deixaram um bilhete para você / they left you a message).

In light of the discussion above, this paper, which is part of an ongoing PhD research project, aims to provide a quantitative account of the strategies of subject indetermination both in Brazilian Portuguese and in English in two representative corpora, The Corpus of Contemporary American English (Davies, 2008-) and Davies and Ferreira’s (2006-) Corpus of Portuguese. The frequency descriptions are meant to display speakers’ preferred strategies of subject indetermination in both languages paving the way for a more robust explanatory account of the reasons why speakers of these two languages use different strategies for the same pragmatic intention.

References


Conventionalized linguistic exponents are frequently found in the discourse of experienced writers of a genre (Cheng, 2011). These linguistic patterns aid the effective communication of meaning, since they are readily recognized by readers who associate them with different moves or steps of the genre in question. The effective instruction of genre should preferably involve the introduction of these specific language features (Henry & Rosenberry, 2001). Doing so would help students better exploit the discourse structures of the register in which they are being initiated (Hyland, 2012). As a result, previous studies have attempted to link different lexico-grammatical features to the move scheme developed by Swales (1981, 1990, 2004). One study that has tried to establish this link is Cortes (2013), who established a connection between a particular linguistic features (i.e., lexical bundles) and Swales’ move scheme in a multi-disciplinary corpus of research article introductions. Her findings are a considerable contribution to move-schema theory through providing a detailed description of the communicative function of moves. Cortes’ (2013) study was based on a multidisciplinary corpus of RA introductions, and since lexical bundles have been found to be discipline-bound (see Cortes, 2004; Cortes, Jones, & Stoller, 2002; Hyland, 2008), a similar investigation into the connection between lexical bundles and move structure in a unified discipline would allow us to gain a more in-depth understanding of the linguistic means through which moves are realized in that specific discipline. It would also greatly contribute to the process of second language academic writing instruction within the scientific domain in question. Therefore, the aim of the present study is to provide readers with a list of the most frequently-recurring lexical bundles in a corpus of RA introductions from the field of applied linguistics. Following the methodology of Cortes (2013), we then sought to examine the relationship between these bundles and the moves and steps in which they were found. We believe that such a list can have great pedagogical applications for EAP instruction.

For this study, a corpus consisting of 1,009 RA introductions from the field of applied linguistics was compiled. The word count stood at one million word tokens. The articles were randomly selected from five high-ranking journals of the field, namely Applied Linguistics, English for Specific Purposes, Journal of English for Academic Purposes, Journal of Second Language Writing and TESOL Quarterly. Lexical bundles of four+ words were identified using WordSmith Tools 7.0 (Scott, 1996). The same cut-off points used by Cortes (2013, p. 36) were also adopted for this study. The final list (excluding bundles that formed parts of longer sequences and those that were discipline specific) included 6342 lexical bundle tokens and 229 types. All bundles in the final list were classified both in terms of their structure and function using the taxonomies introduced by Biber, et al. (1999) and Biber, Conrad and Cortes (2003, 2004).
The bundles were subsequently analyzed in their surrounding context with the aim of exploring the rhetorical moves and steps in which they occurred. The most number of bundles (i.e., types) were found in Move 1 (315 bundles), while the least number were identified in Move 2 (112 bundles); 170 bundle types were also observed in Move 3. Our corpus analysis also showed that some of the identified bundles in our corpus were exclusively used to fulfill the communicative aims of one move or step. These bundles are of particular interest since they are widely used by writers to fulfill the rhetorical aims of the move and step in which they occur. We also found that the longer bundles of our study mostly performed a trigger function in that they were used to initiate or ‘trigger’ a move/step.

In addition, the functional classification of the bundles showed that, in line with previous studies which have looked at the function of lexical bundles in research articles (e.g., Biber, et al., 1999; Biber, 2006), most of the bundles in this study, regardless of the move and step in which they were found, were referential in nature, while stance expressions were the least common functional type of bundle. A comparison between the moves and steps reveals that the most number of stance expressions were observed in Move 1 followed by Moves 2 and 3, respectively. According to Swales (1990), Move 1 aims to highlight the importance of the field of study for the discourse community. It suggests that the rhetorical aims in Move 1, especially Step 2 (i.e., making topic generalizations) and Step 3 (i.e., reviewing items of previous literature) may prompt writers to use more stance expressions.

References


Phrasal Complexity in the Writing of Iranian EFL College-Level Students

Hesamoddin Shahriari, Ahmad Ansarifar and Reza Pishghadam (Ferdowsi University of Mashhad, Iran)

Syntactic complexity has been recognized as an important construct in L2 writing instruction and research. Among the list of measures used to analyze syntactic development and complexity in second language academic writing, phrasal embedding has been largely ignored and has only recently been receiving attention from researchers of the field. Studies into syntactic complexity (e.g., Beers & Nagy, 2007; Ellis & Yuan, 2004; Hunt, 1965; Jiang, 2012; Stockwell & Harrington, 2003) have traditionally focused on clausal embedding and subordination measures based on the assumption that academic writing derives its complexity from the elaborate use of clausal constructions (Biber & Gray, 2011). However, a number of studies (e.g., Biber & Clark, 2002; Biber & Gray, 2011; Biber, Gray & Ponpoopon, 2011) have found academic writing to be characteristically dense with non-clausal phrases and complex noun phrase constructions, while reporting a relative absence of clausal elaboration. As a consequence, more studies are needed to investigate the role of phrasal embedding in the development of complexity in academic writing. Studies such as Lu (2011), Parkinson and Musgrave (2014) and Biber, Gray and Staples (2014) have shown that phrasal features tend to increase across levels of study in second language writing. Ravid and Berman (2010) and Staples, Egbert, Biber, and Gray (2016) have also found that phrasal features develop during university years in the writing of L1 English students.

Biber, Gray and Poonpon (2011) have argued that those seeking to write academic research articles, both native and non-native alike, should become familiar with this characteristic feature of the genre which often involves the frequent use of lengthy noun phrase structures. This could be attributed to the need to convey a large amount of information in a limited number of words. They also showed that complex noun phrases are much more appropriate measures of grammatical complexity in academic writing as opposed to embedded clauses. In line with these findings, Biber et al. (2011), hypothesized a series of stages predicting development in writing complexity through the increased use of complex noun phrase constituents and phrasal constructions. This developmental index involves five stages. According to Biber et al. (2011), these grammatical features can mainly be grouped according to three major grammatical types: finite dependent clauses, nonfinite dependent clauses, and dependent phrases, which can serve three main grammatical functions (i.e., adverbials, complements, and noun modifiers). The present study aims to investigate the development of noun phrase complexity in the academic writing of Iranian graduate students at different levels of academic study, comparing the observed patterns of use to those of expert writers.

To test this hypothesis, we compared three groups of abstracts: one group by MA level L1 Persian writers (using a corpus of 25714 words), a second group by PhD-level L1 Persian writers (with a corpus of 25423 words), and a third by a group of published abstracts by expert writers (with a corpus of 24808 words). In this study, we examined 1. Finite dependent clauses including relative clauses as noun
modifiers, complement clauses controlled by nouns. 2. Nonfinite dependent clauses including, -ing and -ed participles as noun post-modifiers, and preposition + nonfinite complement clauses as post-modifiers. 3. Dependent phrases including, attributive adjectives, participles, nouns as pre-modifiers, possessive nouns, of phrases as noun post-modifiers, other prepositional phrases as noun post-modifiers, adjectives, noun as pre-modifiers, appositives, and multiples prepositional phrases as noun post-modifiers.

These 16 different kinds of noun modifiers in our three sets of data were identified and pre- and post-modifiers were manually coded. As previously noted, since we are examining 16 grammatical features and therefore performing 16 tests simultaneously, we used the Bonferroni correction to avoid spurious positives. This adjusts the alpha value for each comparison to .002 (i.e., .05 /16 = .002). The one-way ANOVA is run to determine whether the normalized frequency for the MA, PhD, and EW groups differ significantly from each other. Our findings showed that out of the 16 types of noun modifiers examined in our study, Iranian graduate students of Applied Linguistics did not differ significantly from expert writers in producing 12 of the categories. Out of four kinds of modification for which significant differences were observed, one could be categorized as nonfinite dependent clauses (-ed participle as post-modifier) and the other three as dependent phrases (premodifying nouns, adjective/ noun combination as pre-modifiers, and multiple prepositional phrases as post-modifier). This is largely consistent with Biber et al.’s (2011) hypothesized developmental stages of syntactic complexity, in which finite dependent clauses are predicted to be acquired at earlier stages of writing development, while nonfinite dependent clauses are said to be mastered at intermediate stages followed by dependent phrases which are predicted to appear in the final stages of academic writing development.

We additionally found that abstracts by PhD-level graduate students better approximated those of expert writers. Of the four categories that turned out to be significantly different across the three groups (i.e., nouns as pre-modifiers, ed participles as post-modifiers, adjective/noun sequences as pre-modifiers, and multiple prepositional phrases as post-modifiers), the PhD group of abstracts, when compared to expert writers, only lacked multiple prepositional phrases as post-modifiers (i.e., the last stage in Biber et al.’s (2011) developmental stages of syntactic complexity). The abstracts by the MA group, on the other hand, fell short in all of the four features when compared to expert writers.

Our results have some useful implications for both L2 writing pedagogy and research. First, awareness of the developmental index of noun phrase complexity, as proposed by Biber et al. (2011), and attention to the favored patterns of noun modification by expert writers could greatly help L2 writing instructors, who stand to benefit from taking these patterns into consideration when teaching and drawing their students’ attention to features that are expected to be learned at each stage of development. Our findings are also beneficial for L2 learners who can gain awareness of features distinguishing their writing from those of expert writers in their field of study. Moreover, the findings could also be of great use to writing assessment experts.
References


Age-related patterns in lexical bundle usage: Evidence from a corpus of vernacular Japanese

Kevin Heffernan (Kwansei Gakuin University, Japan) and Yo Sato (Satsuma Language Services, England)

Studies on cognition and aging report slower cognitive performance with increased age (Deary et al., 2010; Nyberg et al., 1996; Pauls et al., 2013; Peelle et al., 2010). However, we do not fully understand the causes of reduced performance. On one hand, these researchers argue that reduced performance is a consequence of reduced processing speed, which in turn is a consequence of aging of the brain. On the other hand, Ramscar and colleagues argue that reduced processing speed simply reflects lifelong learning (Ramscar et al., 2014). These two arguments make predictions about lexical bundle usage. Motivated by these two stances—the aging hypothesis and the learning hypothesis—we ask the question: Does lexical bundle usage show age-related patterns? We review these stances in more detail, and then posit several hypotheses based on them. We then present our investigation of the relationship between lexical bundle usage and age.

These two stances make different predications about age-related trends in lexical bundle usage. Lexical bundles are frequently occurring sequences of words, such as ‘have a look at’ and ‘don’t you think that’. Due to their high frequency, speakers store lexical bundles as a cohesive package in declarative memory. Declarative knowledge includes our semantic knowledge of language, and our episodic knowledge of facts and events. Declarative knowledge also encompasses idiomatic knowledge of proverbs, collocations, and lexical bundles (Ullman, 2016; van Lancker Sidtis, 2012), although the exact nature of this knowledge is still unclear. Importantly, the high level of cohesion within a lexical bundle leads to its faster processing compared to less frequent expressions (Siyanova-Chanturia et al., 2011).

The advocates of the aging hypothesis argue that our reduced processing performance is a consequence of biological changes. For example, Peelle et al. (2009) examined brain activity in older and younger adults when processing sentences with two levels of syntactical complexity. They found that when processing the complex sentences, older adults produced less neural activity in the areas of the brain associated with syntactic processing, and showed less overall coordinated activity between different areas. Morrison and Baxter (2012) present evidence for similar age-related trends in cognitive activity in both language-speaking humans and non-language-speaking animals such as rhesus monkeys. They claim that “[t]he cognitive processes that are mediated by the hippocampus... are those that are most vulnerable to ageing.” The hippocampus “underlies the rapid linking (binding, associating) of different bits of knowledge or experiences” (Ullman, 2016), that is, our episodic knowledge. These researchers argue that the human brain ages, which has consequences for language usage in old age. In contrast to the aging hypothesis, Ramscar et al. (2014) argue that older adults’ reduced cognitive response speed is not evidence
for cognitive decline. They instead posit that the reduced response patterns are a result of the added volume of information a person acquires through day-to-day experiences. Choosing between or recalling items such as memories or lexical information becomes more difficult as the number of stored items increases. Using computer simulations of adult learners acquisition of vocabulary over the adult lifespan, they demonstrated that as an individual’s experience grows, his or her knowledge will increase, and that this will in turn raise the processing costs in his or her cognitive system. Ramscar et al. (2014, p.34) conclude that “there is no neurobiological evidence for any declines in the processing capacities of healthy older adults.”

If older speakers struggle with the reduced performance of an aged declarative memory, then we expect less utilization of lexical bundles with increased age. On the other hand, if the underlying cause of older speakers’ slower processing is larger volume, then we expect there to be no correlation between age and lexical bundle usage, since by definition lexical bundles are few in number. Another possibility is that older speakers may rely on lexical bundles—which are processed faster—more than younger speakers to compensate for their slower cognitive processing. Furthermore, if increased lexical knowledge due to older speakers’ greater life experience plays a role, then we expect that older speakers will use a greater variety of lexemes. However, we do not expect a corresponding increase in lexical bundles as we assume that adult do not continue to learn lexical bundles throughout their lifetime, again because there is a very small number of them.

We used the Corpus of Kansai Vernacular Japanese, a collection of 152 conversational interviews between students (the interviewers) and family members or close acquaintances (the interviewees). The average length of an interview is 55 minutes. The corpus also contains personal information about each interviewee, such as age and education level. Each interview was transcribed, parsed into morphemes and tagged with part of speech information. We limit our investigation to the interviewee (hereafter, speaker) data, which consists of 885,027 morphemes. We divided the speakers into four age groups: students (N = 49, age range 15 to 23 years old), younger adults (N = 32, age range 24 to 39 years old), middle-aged adults (N = 34, age range 40 to 59 years old), and older adults (N = 37, age range 60 to 79 years old). In order to determine the lexical bundles, we took an equally-sized sample of 171,000 morphemes (the size of the smallest group) from each age group. Following Biber et al. (2004), we compiled a list of the four-morpheme sequences that occurred at least 40 times per million morphemes (or in our case, at least 27 times in our balanced sample). This methodology yielded 155 bundles, of which we removed 12 bundles that we judged to be specific to our data, such as *ni-sun-de-ru* ’I live in.’

We calculated the lexical bundle usage rate for each speaker as the portion of his or her speech contained within lexical bundles. The average rate is 4.2% (SD = 1.78), and ranged from 0.9% to 10.6%. We also calculated the type-token ratio for both lexical bundles and lexical morphemes (common nouns, verbs, and adjectives) for each speaker. The lexical bundle type-token ratios ranged from
23.0% to 95.0%, with an average of 58.0% (SD = 13.8). The speakers produced 381,009 lexical morphemes (42.3% of the data). The lexeme type token ratios ranged from 17.1% to 45.5%, with an average of 29.6% (SD = 5.0).

In order to determine the relationship between age and our three measurements (lexical bundle usage rate, lexical bundle type-token ratio, and lexeme type-token ratio), we conducted three one-way ANOVA tests with a Bonferroni-corrected p value of 0.017. Means (with standard deviations in parenthesis) for each group are presented in Table 1. There are significant differences in the lexical bundle usage rates ($F(3, 148) = 8.39, p < .001$). However, the posthoc Tukey test showed that the significant differences were between the older adults and the other groups (all comparisons $p < .01$). None of the other groups differed significantly with each other. In contrast, the ANOVA tests on the lexical bundle and lexeme type token ratios did not reveal significant differences between any of the groups.

These results support the aging hypothesis more than the learning hypothesis. First, the reduced usage of lexical bundles is consistent with the claim that our declarative memory ages. Second, similar to Peelle et al. (2009), we found between-group differences for the older adults only. The middle-aged adults did not show usage patterns that differed from the students, in spite of the fact that the middle-aged adults have two to three times more life experience than the students. But if our declarative memory ages, then why do the older adults not show reduced type token ratios? In the case of lexical bundles, there are so few of them— compare 143 lexical bundles to the 15,298 lexicemes used by these speakers—that even an aged memory may easily store them all without loss. More importantly, we need to make a distinction between two types of declarative memory: episodic memory and semantic memory. Previous studies have demonstrated that the episodic memory is prone to aging but semantic memory is not (Nyberg et al., 1996). If lexical bundles rely more on episodic memory than on semantic memory, then we are able to account for the seemingly contracting claims of the aging hypothesis and the learning hypothesis. The aging hypothesis applies primarily to linguistic functions in which episodic memory plays an important role, whereas the learning hypothesis applies to linguistic functions in which semantic memory plays and important role. This distinction between episodic memory and semantic memory, and the implications for language processing, needs to be explored in future research.

Table 1: Mean rates and standard deviations for the three measures by age group

<table>
<thead>
<tr>
<th></th>
<th>Students</th>
<th>Younger adults</th>
<th>Middle-aged adults</th>
<th>Older adults</th>
</tr>
</thead>
<tbody>
<tr>
<td>LB rate (%)</td>
<td>4.28 (1.96)</td>
<td>4.24 (1.53)</td>
<td>4.32 (1.51)</td>
<td>2.73 (1.80)</td>
</tr>
<tr>
<td>LB type token rate (%)</td>
<td>55.36 (14.0)</td>
<td>59.42 (12.34)</td>
<td>57.12 (14.91)</td>
<td>60.89 (13.65)</td>
</tr>
<tr>
<td>Lexeme type token ratio (%)</td>
<td>29.71 (3.89)</td>
<td>31.47 (5.61)</td>
<td>29.53 (4.99)</td>
<td>28.07 (5.25)</td>
</tr>
</tbody>
</table>

Note. LB = lexical bundle.
References


In the first part of the paper, we introduce a new bottom-up approach to the identification/extraction of multi-word expressions (MWEs) in corpora. This approach is called MERGE (for Multi-word Expressions from the Recursive Grouping of Elements), and involves the successive merging of bigrams to form word sequences of various lengths. More specifically, the approach involves multiple iterations of looping over a corpus to count all tokens (words and n-grams from previous iterations), compute an association measure for all pairs of tokens (currently the log-likelihood measure but other association measures can also be used), and identifying the highest scoring pair of tokens. This pair then gets merged into a new single token, all frequency statistics in the corpus are updated to reflect the new token due to the merger, and the process reiterates for, say, a user-defined number of iterations. Some crucial advantages of this approach are that (i) MERGE is a thoroughly bottom-up procedure requiring few potentially subjective decisions other than the association measure which is computed (which in turn means it’s flexible and includes more information than just cooccurrence frequency); (ii) MERGE does not just return n-grams of a certain pre-defined n, but can return all sorts of n-grams; and (iii) the method requires only a simple adjustment to deal with discontinuous sequences.

In the second part of the paper, we discuss four different validation studies to test/validate the performance of the proposed MERGE algorithm. First, we applied the algorithm to the combined Santa Barbara Corpus of Spoken American English and ICE-Canada to identify MWEs that, according to MERGE, are 'good' and 'bad' MWEs, which was operationalized on the basis of when during the iterations - early vs. late - a MWE was identified by MERGE; we let MERGE run for 20K iterations and allowed for 1-word gaps in the processing. We then had 20 native speakers of American English (students at UCSB) rate them on a Likert scale for how much the MWE stimuli constituted "a complete unit of vocabulary" to see whether the native speakers distinguished between 'good/early' and 'bad/late' MWEs with their ratings, thereby supporting the MWE-quality ratings implied by MERGE. We analyzed
the ratings with a linear mixed-effects model using the MWEs' lengths (as a polynomial to the 2nd degree) and MERGE's ranking (early vs. late) as fixed-effects independent variables and varying intercepts for all MWEs and varying slopes for length and rank for all MWEs and subjects. We found that MERGE's output indeed distinguishes significantly between 'good' and 'worse' MWEs (t=-19.17, df=31, p<0.0001) such that, as expected, 'good/early' MWEs receive better ratings than 'bad/late' ones, with ratings decreasing as MWEs become longer.

Second, we compared the output of MERGE when applied to the same corpus data as above to the output of Brook O'Donnell's (2011) Adjusted Frequency List, a conceptually similar approach, but one that does not take association strength between elements of MWEs/tokens into consideration. We took the 1000 first items returned by MERGE and the 1000 top items of the AFL and randomly sampled MWEs from them for another rating experiment. In that experiment, 20 (different) native speakers judged altogether 360 items on a Likert scale again. Again we analyzed the ratings with a linear mixed-effects model using MWE method (MERGE vs. AFL) and MWE length as independent variables (again as a polynomial to the 2nd degree) and varying intercepts for subjects as well as varying slopes for MWE method per participant. We found that, as hypothesized, when MWE length is controlled for, the MWEs returned by MERGE score significantly higher than those returned by the AFL (t=2.128, df=23.4, p_{1-tailed}=0.022).

Third, we applied both MERGE and AFL to the complete spoken component of the BNC to determine how well both methods can identify 388 expressions that the compilers of the BNC decided to tag as multi-word units (using the <mw></mw> tag). We took the top 10,000 items from either approach and used one-tailed binomial tests to compare the proportion of BNC multi-word units that either approach would identify, i.e. whether MERGE would perform better or worse than AFL on this task; given the previous results, we expected MERGE to find a higher percentage of multiword units than the AFL. Both approaches find mostly high-frequency rather than low-frequency MWEs but do not appear to have a preference for items with high degrees of dispersion. We did separate one-tailed binomial tests in both directions (against both baselines) and found that (i) MERGE found more BNC multi-word units (28.9%) than the AFL (24%) and (ii) that that difference is significant in either direction (P_{MERGE vs. AFL}=0.01522, P_{AFL vs. MERGE} = 0.0178).

Finally, we explored MERGE's performance using L1-acquisition corpus...
data. We ran MERGE on both the adult and the child utterances of the Lara and Thomas corpora from the CHILDES collection. Specifically, we split adult and child data into an early (the first $\frac{2}{3}$) and a late part (the last $\frac{1}{3}$) and compared the MERGE scores of the adult MWEs that the children used in the late partition to the MERGE scores of the adult MWEs that the children did not use later; more precisely, we computed the proportion of MWEs that the children learned within each of several dozens of bins as defined by log-likelihood scores. We then fit a linear model with the square roots of proportions of learned sequences as the dependent variable (we needed to take square roots to avoid violations of linear modeling assumptions) and we used log-likelihood bin as well as MWE lengths (as a polynomial to the second degree) and which child's data we were exploring as independent variables. We obtained a highly significant and highly explanatory model ($p<0.0001$, adj. $R^2=0.7801$) but, most importantly, we found a significant 3-way interaction of MWE length, MWE strength of association (as defined by log-likelihood bin), and child ($p=0.0076$): For both children (with slight differences between them), MWEs with higher MERGE scores are indeed those that children are more likely to learn even when length is controlled.

We conclude by integrating all results, discussing their implications, and suggesting future analyses.
Towards a Welsh Semantic Tagger: Creating Lexicons for A Resource Poor Language

Scott Piao (Lancaster University, UK), Paul Rayson (Lancaster University, UK), Gareth Watkins (Cardiff University, UK), Dawn Knight (Cardiff University, UK) and Kevin Donnelly (Independent Researcher, UK)

Semantic annotation and analysis is an important part of corpus linguistics and other research areas (Gacitua et al., 2008; Potts and Baker, 2013; Demmen et al., 2015), and semantic tagging tools have facilitated this type of research to be conducted on a large scale. A major tool is the USAS semantic tagger developed at Lancaster University (Rayson et al., 2004), originally designed for English but has been extended to cover more languages, including Italian, Chinese, Spanish, Portuguese etc. (Löfberg et al., 2005; Mudraya et al., 2006; Piao et al., 2015; Piao et al., 2016). The USAS framework employs a lexical semantic classification scheme containing 21 major semantic categories that are further sub-divided into 232 categories. The multilingual capability of the tagger enables multilingual/cross-lingual semantic analysis under this unified scheme. In the CorCenCC Project (Corps Cenedlaethol Cymraeg Cyfoes: The National Corpus of Contemporary Welsh)¹, we are extending the USAS to automatically annotate Welsh language data, particularly the CorCenCC corpus, with the semantic information.

A critical part of the USAS extension is the development of a Welsh semantic lexicon which provides a wide coverage of Welsh vocabulary and a high quality of semantic classification. Various Welsh lexical resources have been drawn on in building the Welsh semantic lexicon. A major such resource is the Eurfa Welsh/English bilingual lexicon developed by Donnelly (2016), and made available under an open license (GPL). This resource contains valuable lexical information about a large number of Welsh words, including lemma forms, part-of-speech (POS) categories, many multi-word expressions (MWEs), and English translations. From this resource, we extracted 136,468 single Welsh words (inflected forms) and classified them into USAS semantic categories via their English translation equivalents and English semantic lexicons to form the basis of the new Welsh semantic lexicon. In addition, the words are mapped to their lemma forms in order to improve text coverage of the lexicon by allowing each lexicon entry to match with all inflectional variants of the same lemma in the text.

In addition, we expanded the Welsh semantic lexicon by manually compiling closed-class word lists, such as prepositions, conjunctions etc., in order to cover the highly frequent closed-class words in the running text. Another important source for the lexicon are Welsh names, including person names and place names, which we have collected from a number of resources, including the Language Technologies Unit of Bangor University, UK and websites including “Behind The Name”, “Think Baby Names”, and “Wales”.² Through these approaches, the Welsh semantic lexicon was expanded to contain 143,290 Welsh words.

¹ For the details of the CorCenCC Project, see project website: http://www.corcencc.org
² Permissions to use their name resources were obtained from the following organisations:
   a) Language Technologies Unit of Bangor University
      (https://www.bangor.ac.uk/canolfanbedwyr/technolegau_iath.php.en),
   b) Behind The Name (http://www.behindthename.com/names/usage/welsh),
   c) Think Baby Names (http://www.thinkbabynames.com/names/1/welsh), and
   d) Wales UK (http://www.walesuk.info/wales.html).
We carried out an initial lexical coverage evaluation of the Welsh semantic lexicon using a gold standard corpus, which was constructed for evaluating corpus tools in the CorCenCC project. The gold corpus consists of around 15,000 words and contains selected materials from four existing corpora: Kwici (Welsh Wikipedia)\(^3\), Kynulliad3 (Welsh Assembly Proceedings)\(^4\), Meddalwedd (software translations)\(^5\), and LER-BIML (a small corpus of 10 multi-domain texts)\(^6\). The first three corpora were stored in databases, so the selection was made for the first two by selecting the first 100 items where the length was between 20 and 40 words, and for the third by selecting the first 100 longest items. For the fourth, two of the texts were chosen. The aim was to give a reasonable selection (between 2,000 and 4,000 words) of text from different domains, different sources, and differing lengths in order to create a balanced and representative test corpus. Once the text was gathered, it was cleaned (for example, HTML tags were removed, as were items that contained little Welsh), and then typos and punctuation errors were corrected. Items which contained English words were retained: since modern, less formal Welsh usually contains some English code switches, it is considered desirable that our part-of-speech (POS) tagging and semantic tagging systems have the capability to handle such noise.

<table>
<thead>
<tr>
<th>Content words</th>
<th>Function words</th>
<th>Person/Place names</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of entries</td>
<td>136,468</td>
<td>264</td>
</tr>
<tr>
<td>Sample lexicon entries</td>
<td></td>
<td></td>
</tr>
<tr>
<td>abacws</td>
<td>Eg N3.1</td>
<td>á</td>
</tr>
<tr>
<td>bri</td>
<td>Eg X9.2+</td>
<td>ag</td>
</tr>
<tr>
<td>chwerthin B</td>
<td>E4.1+/X3.2</td>
<td>amdanat</td>
</tr>
<tr>
<td>defnyddio B</td>
<td>A1.5 S7.1+</td>
<td>atoch</td>
</tr>
<tr>
<td>llwybro B</td>
<td>M1 L2 X9.2+/A12+</td>
<td>chithau</td>
</tr>
<tr>
<td>llwydda B</td>
<td>X9.2+/N4</td>
<td>cyn</td>
</tr>
<tr>
<td>plesiwn B</td>
<td>E4.2+/E2+/X7+</td>
<td>erbyn</td>
</tr>
<tr>
<td>plicio B</td>
<td>A9+/A1.1 F1</td>
<td>fel</td>
</tr>
<tr>
<td>tripio B</td>
<td>M1 M2 S8-</td>
<td>gyda</td>
</tr>
<tr>
<td>wailio B</td>
<td>H2</td>
<td>hebddi</td>
</tr>
<tr>
<td>warws</td>
<td>Eg A9+/H1</td>
<td>imi</td>
</tr>
<tr>
<td></td>
<td></td>
<td>lle</td>
</tr>
<tr>
<td></td>
<td></td>
<td>mai</td>
</tr>
<tr>
<td></td>
<td></td>
<td>na</td>
</tr>
<tr>
<td></td>
<td></td>
<td>oddiar</td>
</tr>
<tr>
<td></td>
<td></td>
<td>wrthych</td>
</tr>
</tbody>
</table>

Table 1: Statistics of Welsh semantic lexicon and sample entries.

In the evaluation, our prototypical Welsh semantic tagger based on the current version of the semantic lexicon covered 72.42% of the words in the gold corpus. If the noise in the text mentioned above is excluded, a higher lexical coverage can be expected. Table 1 shows the sizes of three main types of the Welsh semantic lexicon entries and some sample entries from each type. As shown in Table 1, the first column in each lexicon entry contains a word/lemma, the second column contains a part-of-speech tag, and the last column contains possible USAS semantic tag/s (for definitions of the USAS tags, see website: http://ucrel.lancs.ac.uk/usas). The tags contained in the sample entries are from a new Welsh POS tagset developed in the CorCenCC Project, which are defined as follows:

---

\(^3\) See website http://cymraeg.org.uk/kwici.
\(^4\) See website: http://cymraeg.org.uk/kynulliad3
\(^5\) See website http://techiaith.cymru/corpws/Moses/Meddalwedd
\(^6\) See website  http://www.lancaster.ac.uk/fass/projects/biml
We are further expanding the Welsh semantic lexicon in order to achieve a higher lexical coverage and better quality in the semantic classification of the words. For example, we have extracted additional large Welsh word lists from a number of existing Welsh corpora, including *CEG Cronfa Electroneg o Gymraeg* (Ellis et al., 2001), *Kwici* (Corpus of Welsh Wikipedia http://cy.wikipedia.org) and *Corpus of Children's Literature in Welsh* (http://www.egni.org). When we repeated the lexical coverage evaluation including the additional word list, our Welsh lexicons covered over 97% of the text of the gold test corpus. Note that the additional raw Welsh word collection has not yet been built into the semantic lexicon, but such a high lexical coverage figure shows the potential of our semantic tagger that can be achieved when a major part, if not all, of the words already extracted are integrated into the formal semantic lexicon. A prototype of the semantic tagger tool has been built based on the existing Welsh semantic lexicon for testing, which is available at website http://phlox.lancs.ac.uk/ucrel/semtagger/welsh and the Welsh lexicon is available to download under a Creative Commons licence at https://github.com/UCREL/Multilingual-USAS. The semantic tagger will be continuously improved during the project, and we will provide a demonstration of the current version tool.

The current Welsh semantic tagger is at an early stage of development. As the CorCenCC project progresses, we will continue to expand and refine the Welsh semantic lexicons and improve the semantic tagger for annotating Welsh corpus with a high accuracy. In future work, we will: 1) further expand the size and improve quality of the semantic classification of the Welsh lexicon entries, 2) build Welsh multiword expression semantic lexicon, and 3) develop an efficient Welsh semantic annotation tool by combining these semantic lexical resources and word sense disambiguation methods with the Welsh part-of-speech tagger and lemmatiser being created in the CorCenCC project.

**Acknowledgement**

The research on which this article is based is funded by the UK Economic and Social Research Council (ESRC) and Arts and Humanities Research Council (AHRC) as part of the *Corpws Cenedlaethol Cymraeg Cyfoes (The National Corpus of Contemporary Welsh)* (Grant Number ES/M011348/1).

**References**


The textual colligation of stance phraseology in cross-disciplinary academic discourses: the timing of authors’ self-projection

Jihua Dong and Louisa Buckingham (University of Auckland, New Zealand)

Lexical items, according to Hoey (2005, p.13) “are primed to occur in or avoid, certain positions within the discourse”. An analysis of textual colligation, the term Hoey (2005) uses to denote such priming, explores the textual position of linguistic markers in relation to textual structures. Recent studies have examined the textual colligation of particular words or phrases (e.g., Hoey & O’Donnell, 2008; Mahlberg, 2009; O’Donnell et al., 2012). Textual colligation explores the textual position of linguistic markers in relation to textual structures and the interaction between the textual position and discourse functions (Hoey, 2005). Previous studies have enriched our understanding of textual colligation of particular linguistic features such as keywords or key phrases in a text. This study investigates the textual colligation of a type of linguistic marker typical for one particular semantic group, namely, stance. Informed by previous approaches to stance (e.g., Biber et al., 1999; Hunston, 2011; Hyland, 2005; Martin & White, 2005), the study employs a classification framework which distinguishes between four types of stance phrases: attitude, reference, hedges, and cognitive (or epistemic).

This quantitative study investigates the textual colligation of stance phrases in academic discourse in the disciplines of agriculture and economics. The study employs a purpose-built corpus of 655 published research articles totalling around 3 million tokens. We use Wordskew software (Barlow, 2016) to investigate the position (or colligation) of stance phrases at the level of sentence, paragraph and text, and examine the existence of disciplinary variation with respect to the textual colligation of these phrases.

Disciplinary variation was found in the association between the stance phrases and the particular text positions (sentence, paragraph and text), which may be a reflection of disciplinary practice. For instance, the agriculture corpus contained significantly more cognitive, attitude and hedge stance phrases in the sentence-initial and medial positions. In contrast, the economics corpus contained significantly more reference stance phrases in the sentence-initial and medial position, reflecting perhaps the disciplinary practice in economics to depart from making reference to a text (e.g., to the author’s own study, an external resource, shared knowledge or a different part of the same text), when presenting a proposition.

Nevertheless, the proportion of stance phrases in each of the three textual positions is notably similar in the two disciplines. It may be inferred that the textual position of particular stance phrases may be a result of the type of routinized discourse or communicative function these serve (Hoey, 2005). Reference phrases have a tendency to occur in the sentence-initial positions, possibly due to the fact that the sentence-initial positions allow writers to treat these phrases as the point of departure for new information in the sentence. Cognitive, hedges, and attitude phrases were found to have a preference for the sentence-medial positions. The association suggests that the researchers in the two disciplines...
tend to express their projection in a less notable way by expressing stance in sentential non-salient positions. In contrast, a smaller proportion of stance phrases occurred in sentence-final positions. At the level of paragraph and text, both disciplines tend to use a high percentage of stance phrases in the medial position. Hedge phrases, however, occurred more frequently at the paragraph- and text-final than the initial position.

The use of Wordskew has contributed to revealing the text positions at the sentence, paragraph, and text level. It provides an efficient way to quantify the textual position of particular linguistic features, and contributes to visualising the distribution of particular linguistic features in the organization of a text.

References


Assessing the impact of an intensive EAP programme: multidimensional analysis of an authentic learner corpus
Steve Issitt (University of Birmingham, UK)

Multidimensional analysis (MDA) forms a central component of the paper. MDA uses computers to count the frequencies of linguistic features in tagged corpora, and then to carry out a factor analysis on these feature counts in order to identify ‘dimensions’, that is, sets of meaningful associations among individual variables.

The multidimensional approach can be used to identify individual features which can then be quantified and compared so that it is possible to examine a particular text, identify a feature, for example the incidence of agentless passives or average word length and compare it with another text. This enables an analysis of pre and post-writing samples which the study presents as a key methodological technique.

The current study applies multidimensional analysis to a learner corpus, namely the EAPCORP which stands for English for Academic Purposes Corpus and consists of 526 individual scripts and 263 matched pairs (pre and post course) representing a complete set of essays for two separate EAP summer programmes. All the scripts were written under exam conditions by hand and copied in their entirety with no corrections. Students were given a thirty minute writing task at the beginning of their respective EAP programme which were of twenty, fifteen, ten or six weeks duration. They were given the same writing task at the end of their programmes. The task was "What are the differences between your country and the UK?". The question was chosen because of its simplicity, relevance and the fact that it offers considerable opportunity for explanation, comparison and exposition. The aim was to standardise the prompt question without creating too much boredom and to prevent students from preparing an answer in advance.

The corpus itself presents data for two EAP programmes with two separate student cohorts, one from 2010 and one from 2012, offering an opportunity to examine developmental features. The EAPCORP itself may have some advantages over some other learner corpora in that it is designed to measure linguistic feature movement for a specific programme and can be considered in this sense to be purpose-built. It also covers two separate years and has a relatively large number of matched pair samples.

The data was analysed by using the multidimensional analysis tagger (MAT) which "replicates Biber's (1998) tagger for the multidimensional functional analysis of English texts ... and generates a grammatically-annotated version of the corpus or text selected [and] the statistics needed to perform a text-type or genre analysis" (Nini, A., 2014, p.1). The MAT programme operates by employing the algorithms used by Biber (1988) and employs an adaption of the Stanford tagger (Toutanova, K. et al., 2003) to analyse the text. The programme offers a range of analytical categories of which two are used in the present study, the individual variables and the dimensions.

The results of the study suggest that the students’ written production is marked by movement along a talk to writing strand suggested by Shaw and Liu (1998) and that one characteristic of developing student second language writing, is
text of a more recognisably “written” and to a certain degree more complex character. This is evidenced by the MAT programme analysis which has shown that in terms of individual features there has been an increase in use of longer words, passives, nominalisations, verb participles, sentence relatives, demonstratives, prepositional phrases and attributive adjectives. There has been a decrease in first and second person pronouns, private verbs, use of the present tense, use of the past tense and fewer emphatics. In terms of dimensional characteristics, this movement from talk to writing is further evidenced by the observed shift from nonabstract to abstract information which characterises all eight programmes (2010 and 2012) and the change from involved to informational production, from high overt to low overt persuasion and from low to high information elaboration which characterise seven out of eight of the programmes.

References

A Corpus-based Analysis of Syntactic Complexity measures in the Academic Writing of EFL, ESL, and Native English Master’s Students
Maryam Nasseri (University of Birmingham, UK)

'Syntactic Complexity' together with 'Lexical Complexity' are frequently regarded as linguistic subdomains or subsystems related to the concept of 'Linguistic Complexity' (Kuiken, et al., 2010; Szmrecsanyi and Kortmann, 2012) and are used to analyse linguistic performance, proficiency, and development in first and second language acquisition research studies. They are oftentimes linked to the studies which describe or analyse Complexity, Accuracy, and Fluency (CAF measures) as initially proposed by Skehan (1989) and continued in the works of Wolfe-Quintero, Inagaki and Kim (1998), Wigglesworth and Storch (2009), Norris and Ortega (2009), and Tavakoli and Rezazadeh (2014) among others.

In second language writing studies, however, syntactic complexity measures are taken as variables to investigate the roles of task complexity, writing ability, genre, and teaching methods across different proficiency levels, age groups, and developmental periods (Foster and Skehan, 1999; Chandler, 2003; Ellis and Yuan, 2004; Lu, 2011; Ahmadian, 2012; Wu and Ortega, 2013) as well as the impact of pedagogical intervention on the development of linguistic features (Ortega, 2003). Syntactic complexity measures are suggested for use in L2 development bands as placement criteria (Larsen-Freeman, 1978) and reported to be good indicators of the overall development in second language (Arthur, 1979; Ferris, 1994; Lu, 2010), proficiency levels (Wolfe-Quintero, Inagaki, and Kim, 1998), and specifically writing ability (Rafoth and Combs, 1983).

A number of measures have been proposed for assessing syntactic complexity, among which Mean Length of T-unit, Mean Length of Clause, Mean Number of Clauses per T-unit, and Mean Number of Dependent Clauses per Clause are reported to be good indicators of proficiency levels in second language writing (Cooper, 1976; Witte and Davis, 1982; Wolfe-Quintero et al., 1998; Ortega, 2003; Norris and Ortega, 2009; Lu, 2010). Some experts developed coordination and subordination indices and found out that coordination values are higher in beginner learners’ writing and subordination values are higher in advanced-level learners’ production (Monroe, 1975; Bardovi-Harlig & Bofman, 1989; Lu, 2010).

Many of these studies, however, have focused on only one or a few syntactic measures as determiners of students’ performance or proficiency levels, and this has often led to mixed or contradictory results. The present study includes an extended list of syntactic complexity measures based on the Lu and Ai (2015) classification, providing more grounds for the actual differences to be spotted among the participants. What follows is an account of an analysis of such measures in the English academic writing of three groups of university graduates with different first languages in order to detect possible differences in the use and the pattern of various syntactic structures as indices of syntactic proficiency.
Research Questions and Methodology

In the second language and academic writing, the comparative studies which analyse the complexity of syntactic structures of texts produced by non-native vs. native English students, have been of special interest to researchers both in terms of the performance and the proficiency level (Ai and Lu, 2013; Qi, 2014; Lu and Ai, 2015). To find out whether the findings of this study support or reject the previous research studies’ results in respect of the differences in English academic texts from L1 and L2 writers, and concerning the Abstract section as a distinct sub-genre of a dissertation, the following research questions are formulated:

1. Are there any significant differences between EFL and ESL students in any of the fourteen measures of syntactic complexity produced in the Abstract section of master’s dissertations?
2. Are there any significant differences between EFL and English NS students in any of the fourteen measures of syntactic complexity produced in the Abstract section of master’s dissertations?
3. Are there any significant differences between ESL and English NS students in any of the fourteen measures of syntactic complexity produced in the Abstract section of master’s dissertations?

Method, Data Analysis, and Findings

This study analyses the Abstract section of master’s dissertations in Applied Linguistics and any other EFL-related disciplines written by EFL (English as a Foreign Language), ESL (English as a Second Language), and NS (Native Speakers of English) students in terms of syntactic complexity. The EFL students are all Iranian master’s students with varying L1s who studied in various universities in Iran with a centralised curriculum and submitted their dissertations within five years prior to the commencement of this study. The ESL and NS students are all master’s students who studied and submitted their dissertations in various universities in the UK; the ESL students are from different nationalities and language backgrounds.

The corpus is drawn from a homogeneous age group (20-40 years old) of female and male students; 50 abstracts from each group respectively (a total of 150 abstracts). The corpus was analysed using the Syntactic Complexity Analyzer (L2SCA), a computational system for the automatic analysis of syntactic complexity developed and reliability-tested by Lu (2010, 2011).

The frequencies of nine syntactic structures as production units are reported and, based on that, fourteen measures of syntactic complexity, under five syntactic categories based on Lu (2015) classification are analysed:

A. Length of Production Unit: MLS (Mean Length of Sentence), MLC (Mean Length of Clause), and MLT (Mean Length of T-unit)
B. Sentence Complexity: C/S (Sentence Complexity Ratio)
C. Amount of Subordination: C/T (T-unit Complexity Ratio), CT/T (Complex T-unit Ratio), DC/C (Dependent Clause Ratio), and DC/T (Dependent Clause per T-unit)
D. Amount of Coordination: CP/C (Coordinate Phrases per Clause), CP/T (Coordinate Phrases per T-unit), and T/S (Sentence Coordination Ratio)

E. Degree of Phrasal Sophistication: CN/C (Complex Nominals per Clause), CN/T (Complex Nominals per T-unit), and VP/T (Verb Phrases per T-unit)

The respective values were first subject to the Shapiro-Wilk Normality Test and then analysed via 14 one-way ANOVAs, to find any overall differences between the three groups’ production of the mentioned syntactic measures. The Post-hoc analysis then was followed for the measures which showed between-group differences in the ANOVA tests.

ANOVA tests with the threshold level of $p = 0.05$ for six measures of MLC, C/S, T/S, CP/T, CP/C, and CN/C did not show overall significant differences among the groups. The remaining eight measures of syntactic complexity showed significant differences between at least one pairwise comparison among the three groups. Table 1 demonstrates the results of three sets of post-hoc multiple comparison tests for these eight measures. In this study, a difference is considered statistically significant if all three methods confirm the significance of the difference between the pairwise comparisons.

Table 1. P-values of three post-hoc multiple comparison tests for eight syntactic complexity measures across EFL, ESL, and NS groups

<table>
<thead>
<tr>
<th>Measure</th>
<th>EFL vs. NS</th>
<th>ESL vs. NS</th>
<th>EFL vs. ESL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MLS</strong></td>
<td>HSD = 0.02</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Scheffé = 0.03</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>B &amp; H* = 0.02</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>MLT</strong></td>
<td>HSD = 0.00</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Scheffé = 0.00</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>B &amp; H* = 0.00</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>VP/T</strong></td>
<td>HSD = 0.00</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Scheffé = 0.00</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>B &amp; H = 0.00</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>C/T</strong></td>
<td>HSD = 0.00</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Scheffé = 0.00</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>B &amp; H = 0.00</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>DC/C</strong></td>
<td>HSD = 0.00</td>
<td>-</td>
<td>HSD = 0.04</td>
</tr>
<tr>
<td></td>
<td>Scheffé = 0.00</td>
<td>-</td>
<td>B &amp; H = 0.04</td>
</tr>
<tr>
<td></td>
<td>B &amp; H = 0.00</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>DC/T</strong></td>
<td>HSD = 0.00</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Scheffé = 0.00</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>B &amp; H = 0.00</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>CT/T</strong></td>
<td>HSD = 0.00</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Scheffé = 0.00</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>B &amp; H = 0.00</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>CN/T</strong></td>
<td>HSD = 0.00</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
The findings tabulated in table 1 show that the EFL group produced significantly shorter sentences (MLS) and shorter T-units (MLT) than the NS group, supporting the results of Ai and Lu (2013) and Foster and Tavakoli (2009). The amount of subordination as manifested in the four syntactic structures of C/T, DC/C, DC/T, and CT/T was also proved to be significantly lower in the EFL group compared with the NS group. Similar trends were found in the studies of Ai and Lu (2013) and Foster and Tavakoli (2009) among the NNS and NS groups for such measures of syntactic subordination. Tukey and Bonferroni & Holm tests also detected a statistical difference between the EFL and ESL groups regarding dependent clause ratio measure in this category. Finally, concerning the degree of phrasal sophistication, the EFL group produced far fewer verb phrases and complex nominals per T-unit (VP/T and CN/T) than their NS peers which is again consistent with the results of Ai and Lu (2013) where NNS produced a smaller proportion of complex nominals than the NS group.

The p-values obtained from three sets of post-hoc tests confirm that for the above eight measures making the three broad categories of length of production unit (MLS, and MLT), amount of subordination (C/T, DC/C, DC/T, and CT/T) as well as degree of phrasal sophistication (VP/T and CN/T), the EFL group’s performance was considerably lower than the NS group, suggesting that the Iranian Master’s students production of these syntactic structures need to be markedly improved via relevant academic writing courses. The NS group proved to be syntactically more proficient than both NNS groups overall and the EFL group was the least syntactically proficient group. The results could benefit syllabus designers and curriculum decision-makers of EFL-related disciplines in EFL settings to attend to the syntactic proficiency of graduate students to approximate their proficiency level to that of native English speakers, especially in thesis and dissertation writing.

References


The British National Corpus Revisited: Developing parameters for Written BNC2014
Abi Hawtin (Lancaster University, UK)

1. The British National Corpus 2014 project

The ESRC Centre for Corpus Approaches to Social Science (CASS) at Lancaster University and Cambridge University Press are working together to create a new, publicly accessible corpus of contemporary British English. The corpus will be a successor to the British National Corpus, created in the early 1990s (BNC1994\(^1\)). The British National Corpus 2014 (BNC2014) will be of the same order of magnitude as BNC1994 (100 million words). It is currently projected that the corpus will reach completion in mid-2018. Creation of the spoken sub-section of BNC2014 is in its final stages; this paper focuses on the justification and development of the written sub-section of BNC2014.

2. Justification for BNC2014

There are now many web-crawled corpora of English, widely available to researchers, which contain far more data than will be included in Written BNC2014. For example, the English TenTen corpus (enTenTen) is a web-crawled corpus of Written English which currently contains approximately 19 billion words (Jakubíček et al., 2013). Web-crawling processes mean that this huge amount of data can be collected very quickly; Jakubíček et al. (2013: 125) note that “For a language where there is plenty of material available, we can gather, clean and de-duplicate a billion words a day.” So, with extremely large and quickly-created web-crawled corpora now becoming commonplace in corpus linguistics, the question might well be asked why a new, 100 million word corpus of Written British English is needed.

2.1 The enduring popularity of BNC1994

One answer to that question is the demonstrable value even two decades on of BNC2014’s earlier counterpart, BNC1994. Despite being created in the 1990s and containing data from as far back as the 1960s, BNC1994 is still extremely widely used in linguistic research. This is perhaps surprising, because BNC1994 no longer represents contemporary British English, and is certainly not anywhere near the largest available corpus of British English.

A simple search for the term “British National Corpus BNC” in Lancaster University’s online library catalogue yields 935 results (although this figure does include some repeats). Just under half of these results were works published from 2010 onwards. This shows that BNC1994 continues to be a very productive data source for research right up to the present day.

---

\(^1\) The name ‘BNC1994’ is not widely used, but the decision has been made to refer to the corpus in this way in order to make the link between BNC1994 and BNC2014 clear.
So, if BNC1994 is not the largest available British English corpus, and is also not the most contemporary, there must be a different reason why BNC1994 continues to be so productive for linguistic research. To appreciate this, it is informative to revisit some of the stated goals of the creators of BNC1994:

- To create a corpus an order of magnitude larger than any currently freely available corpus.
- To create a synchronic corpus.
- To include a range of samples from the full range of both spoken and written British English.
- To create the corpus using a non-opportunistic design.
  (Burnard, 2002: 53).

I argue that the reason that no large, web-crawled, or more contemporary corpora have enjoyed the level of uptake of BNC1994 is because none of them meet all of these goals in the way that BNC1994 does. Many corpora, such as BE06 (Baker, 2009), are more contemporary than BNC1994 but are much smaller, whilst many other corpora, such as enTenTen (Jakubíček et al., 2013), are more contemporary and much larger, but do not meet the goal of being synchronic, cannot guarantee the language contained in them is British English, and also do not contain samples from the full range of British English.

2.2 The benefits of a ‘hand-made’ corpus

We see then, certain benefits which BNC1994 has over large web-crawled corpora; let us therefore consider in more detail specific examples of how the ‘hand-made’ nature of Written BNC2014 will set it apart from other corpora. By ‘hand-made’, I mean a corpus where texts are selected by manual procedures, perhaps assisted by automatic measures, but with the choice of texts ultimately made by a human and not by unsupervised software as is normally the case for general-purpose web-crawled corpora.

One of the benefits of a ‘hand-made’ corpus is that this method allows the corpus creators to have much greater control over what texts are included, thus guaranteeing with a rather higher level of certainty than in web-crawled corpora that the selected texts were written by speakers of British English. Jakubíček et al. (2013: 126) trained a classifier to distinguish between British and American English when creating enTenTen; such an approach, however, is less rigorous than a human manually ascertaining this information. Of course, to accomplish this effectively for a very large number of texts (approximately 20,000) appropriate procedures and strategies are required, the development of which is a present focus of the BNC2014 project effort.

A further benefit of a ‘hand-made’ corpus is that we will be able to include within BNC2014 data types which a web-crawl cannot – chiefly, published books. A significant amount of time will be dedicated to negotiating with publishers to include published books within Written BNC2014, which will set the corpus apart from many other contemporary corpora. It is planned that Written BNC2014 will contain 41 million words of published books.
Of course this time spent manually collecting certain texts will result in the need to partly automate large areas of data collection in order to assist the manual effort. Our approach to periodicals in particular is to perform mass-downloads from the web using crawling techniques, but then, critically, for project personnel to work manually with this data, and thereby to select appropriate texts for inclusion.

Given the time investment which will be required to seek publisher’s permission to sample extracts of published books, we have determined that for all other kinds of writing, we will not attempt to seek copyright holders’ permission, but will instead take advantage of certain relevant exceptions in UK copyright law (UK Copyright Service, 2015). Written BNC2014 will be a non-commercial project, thus, under the ‘Non-commercial research’ exception, we will not breach any intellectual copyright law in “copy[ing] limited extracts of works” (UK Government, 2014). Such use must be within ‘fair dealing’ and must not lead to any financial impact on the copyright holder (UK Government, 2014). On the latter point, we consider it is highly unlikely that there would be any financial impact on any of the copyright holders of works included in Written BNC2014, because the eventual texts will be so heavily transformed, with XML markup and word-level annotation, that it is doubtful that anyone would try to read the text in Written BNC2014 rather than the original. ‘Fair dealing’, meanwhile, is “a legal term used to establish whether a use of copyright material is lawful or whether it infringes copyright” (UK Government, 2014). Fair dealing is determined on a case-by-case basis, and factors which have been deemed relevant by courts in determining fair dealing include whether the use of the work affects the market for the original, and whether the amount of the work used was reasonable, appropriate, and necessary (UK Government, 2014). The use of works in Written BNC2014 is highly unlikely to affect the market for the original, and, as only small samples (of around 5,000 words) will be taken from texts, this use is likely to be considered to fall within the limits of fair dealing. Approaching data collection under these exceptions to UK copyright law will give the benefit of a corpus which contains a broad mix of carefully selected texts which we could not have achieved without using these exceptions.

3. Conclusion

In conclusion, it is very likely that the British National Corpus 2014 will prove as valuable to the research community as its predecessor, because it will have features simply not present in large web-crawled corpora. This is not to say that web-crawled corpora are not themselves valuable resources for particular purposes, most notably purposes for which the combination of size and speed of collection is especially desirable. A corpus of the size that will take the BNC2014 project team years to create could be created by a web-crawler in a matter of hours. However, smaller ‘hand-made’ corpora such as BNC1994 and BNC2014 have in their own sphere equally many advantages; as this paper has illustrated, our design decisions and procedures for corpus construction have been defined with these particular strengths of BNC-style corpora in mind.

2 This is in stark contrast to BNC1994 where texts were often tens of thousands of words in length, and which contains only 4049 texts.
References


The Developmental Relationship between Spoken and Written Clause Packaging in an English Secondary School
Mark Brenchley (University of Exeter, UK)

This poster will detail the main findings of a new study into the relationship between speech and writing during the secondary phase of the English education system, situating these findings within the context of other recent studies into the syntactic development of L1 speakers and writers during the school years.

The present findings stem from the analysis of a bespoke corpus of 180 pairs of spoken and written non-narratives, with the author directly eliciting each pair from students attending a mainstream secondary school in Southern England. The corpus was further designed so as to be balanced across two core developmental axes: (a) the year group of the student, and (b) their National Curriculum attainment level.

So constructed, the corpus was analysed in terms of the students' modality-related distributions of clause packaging, defined here as comprising the various means by which clauses are combined via coordination and subordination (cf. Berman & Slobin, 1994). To help ensure a more in-depth analysis, an extended set of measures was employed, ranging from the general (e.g. the number of clauses per t-unit and the number of words per clause) to the more specific (e.g. the number of non-finite adverbial clauses per t-unit and per clause, the number of relative clauses per t-unit and per clause, and the number of phrasal clauses per t-unit and per clause).

So analysed, the study indicates adolescent students at the present age and attainment levels to be at a stage where they are capable of differentiating their modality-related syntax, at least for these texts and measures. It also found this differentiation to be something that varied according to the particular kind of packaging measured. Thus, the spoken texts exhibited a greater number of t-units per t-unit complex, together with a greater prominence of finite adverbial and post-verbal complement clauses. Conversely, the written texts exhibited a greater overall prominence of non-finite clauses, whilst neither modality was distinguishable in terms of their respective proportions of relative clauses. Finally, this differentiation was found to be developmentally static, with the students handling their spoken and written clauses in much the same way regardless of their age or attainment level.

Overall, these findings are interpretable in terms of the participants tapping into the differential production conditions of speech and writing, but without necessarily fully exploiting these conditions (Berman, 2008; Biber, 1988, 1992). Moreover, when placed in the context of the wider evidence base (Berman, 2008; Myhill, 2008; Nippold, 2007; Nippold & Scott, 2010; Ravid & Tolchinsky, 2002), the findings suggest two further conclusions. Firstly, they indicate students at the present age and attainment levels to be at a stage where their syntactic output is more in line with the discourse of mature speakers and writers. Secondly, they indicate modality to be an
aspect of student syntax that is potentially characterised by a nuanced sensitivity to the various communicative features of the wider discourse context.

References


Corpus analysis of workplace discourse: The case of the construction industry
Michael Handford (Cardiff University, UK)

This paper will discuss how spoken workplace discourse\(^1\) can be analysed using corpus tools, while arguing that corpus software tools alone cannot provide sufficient insight into workplace interactions to answer many of the important questions about meaning in context. It argues that, along with keywords, collocations, concordance lines and discourse prosody, a further 'tool' that needs to be used in corpus analysis of workplace discourse (CAWD) is the 'understanding context' tool (Handford, 2017). This can be achieved through triangulating with other data sources, and building a plausible interpretation of the interaction in question.

The data analysed here is the 'CONIC' Construction Industry Corpus (JSPS Grant No. 25370423), a corpus of over 350,000\(^2\) words of largely professional English as a Lingua Franca spoken interactions recorded in Singapore, Vietnam, Bangladesh, Hong Kong, Japan and the UK, and supplemented with a wealth of other data. It is these other data that facilitate a plausible contextual analysis of the interactions, including at the co-textual, non-verbal, situational and wider sociocultural levels. These data include several hours of onsite and meeting video data transcribed with ELAN, over 20 hours of interviews with more than 25 professional engineers and other expert informants, observation notes, as well as several written documents (e.g. work plans, meeting agendas, presentation slides and written contracts), workplace images (e.g. signs, photographs and diagrams), recordings of focus group discussions with professionals, and training materials. It will be argued that the collection of such data is a requisite of doing CAWD.

The talk draws on Candlin’s exhortation that, when doing professional discourse analysis, we should seek to move beyond mere description of the data to a more explanatory position. Analysis thus moves beyond describing what, to making plausible explanations of how and why: 'Any analysis of text which aspires to some explanatory rather than merely descriptive adequacy presupposes an engagement with social action within the context of the institution in question, and needs to take account of the distinctive perspectives of the involved participants (including the researcher)' (Candlin, 2006: 6). In other words, we need access to scales of context that may not be evident in the co-text if we are to build on textual patterns, and we need to provide emic accounts of communicative action.

The exploration of the relationship between text and context is a crucially important element of the analysis of professional discourse, because without background knowledge it is often impossible to understand what the interaction or document is about (Charles and Charles, 1999). While the

\(^1\) Workplace discourse is used here as an umbrella term, capturing business, institutional and professional interactions (Koester, 2010).

\(^2\) As such, the corpus may be the largest corpus of ELF professional interactions to date, and is a relatively substantial corpus of professional spoken discourse.
relationship between text and context is, of course, central in all types of discourse analysis (Gee, 2005; Flowerdew, 2014)\(^3\), it is particularly so in professional and workplace contexts because of the high degree of shared knowledge and practices in workplace communities; linguistically, there is a prevalence of both jargon and highly specific references on the one hand and vague or explicitly deictic language on the other – both of which present challenges for the analyst. This is the reason why specialised corpora are potentially more suitable for research into professional discourse than large corpora.

Extract 1 provides an example of the importance of context in understanding professional or workplace texts, and touches of some of the findings to be discussed in the talk.

**Extract 1**

Speaker 1: And then and (I go to this here) they can connect + here. . .
Speaker 2: Hmm
Speaker 1: . . .with (an opposite) side. . .
Speaker 2: Hmm
Speaker 1: . . .and then and then we measure there
Speaker 2: Hmm
Speaker 1: and there
(Handford and Matous, 2011: 96).

Without any contextual information it is impossible to make much sense of this interaction, or why it might be worth analysing. In fact, it is taken from a large construction project in Hong Kong, and speaker one is a male Japanese engineer in his early 30s responsible for a part of the project. He is explaining a procedure that needs following to the foreman, a Hong Kongese male in his early 50s, as the foreman needs to communicate this to the local subcontracted labourers. The interaction takes place in the Portacabin on the site. There is a high degree of place deixis present (e.g. ‘here’ and ‘there’), and the analysis of video data shows that speaker 1 is drawing a picture with a pencil to demonstrate the procedure. Interviews with the engineer and foreman shed further light on the motivations for this interaction, in other words an explanatory level of analysis, such as the higher probability of cooperation from the sub-contracted labourers if the foreman issue such directives, rather than the engineer. The engineer, in contrast to the foreman and the labourers, is a permanent employee of the contractors, which has implications for the way he and his actions are perceived. As such, the interaction is interesting in the way power is negotiated in workplace interactions: while the engineer is more senior in terms of status, he is reliant on the foreman for the completion of this task.

In order to explore the issues discussed above, the following questions will be explored in this talk:

---

\(^3\) Discourse analysis is used here as an umbrella term for any type of analysis, including corpus linguistics, that seeks to understand how people create and interpret meanings in actual situations.
1. What lexicogrammatical features are statistically significant in the CONIC corpus, and how do these patterns compare to other professional contexts?
2. How can we interpret, and not merely describe, such patterns, and what do they imply about the construction-industry context?
3. What might some of the implications from the above question be for training in the industry?

The talk will show how deictic features are statistically significant (Scott, 2011) in construction discourse (Emmitt and Gorse, 2003; Handford, 2014) when compared to everyday discourse \(^4\), unlike many other professional contexts. Through analysis of video data, the types of non-verbal communication used in conjunction with such items will be outlined, along with insights from interview data and other sources as offering explanations as to why such patterns are found. It will also be shown how such an approach allows for insights into other discursive practices, as well as the arguably different expectations towards seemingly face-threatening language (Tsuchiya and Handford, 2014), and the considerable communicative, interpersonal and intercultural challenges specific to this industry (see also Tijhuis and Fellows, 2012; Handford 2014).

References


\(^4\) Using the VOICE corpus of ELF interactions, and also a sub-corpus of the CANCODE corpus as reference corpora.

’I will force myself to believe that everyone is just another monster from doom’: A corpus stylistic analysis of agency in pre-massacre narratives

Emily Powell (Cardiff University, UK)

The language of culpability in criminal accounts elucidates the moral struggle of offenders and may indicate the likelihood of them offending again. While agency has been analysed in offenders’ accounts of past crimes (O’Connor, 2000; Presser, 2004), little research has been carried out into how criminals navigate their responsibility in texts written preceding crime. This paper aims to offer insights into transitions in the agency of criminals before they commit crimes and challenge existing assumptions concerning the link between a sense of agency and transgression.

In the criminal justice system, offenders are encouraged to take responsibility for their actions and acknowledge the impact of their offences on others. This is based on the premise that taking responsibility will encourage desistance and change behaviour. O’Connor (2000) suggests that using more agentive language encourages an offender to take responsibility and may reduce the chance of them reoffending. However, there is some evidence (e.g. McKendy, 2006) that offenders need to explore their own victim role before they can change their behaviour. An analysis of the language of offenders in the midst of their preparations before a crime therefore offers us a unique insight into how they position themselves and others in relation to their current and future actions, without the responsibility imposed on them by the criminal justice system or the benefit of hindsight, and contributes to the debate around the link between responsibility and desistance.

Agentive language has been discussed widely in relation to the use of the passive or nominalisations (Fairclough, 2001), disengaged lexis and pronouns (O’Connor, 2000), reported speech (Schiffrin, 1996) and representational choices such as the objectivation of actions and the deagentialization of actors (van Leeuwen, 1996). However, measuring levels of agency is problematic (Ahearn, 2001) because it assumes that agency is dichotomous and also implies that we can be certain of the reasons for the use of a certain structure or phrase. Alternatively, exploring categories of agency allows us to examine how offenders present their crimes and the different elements of it (e.g. the victims, the condemners, the agent, the weapon), and what they do with their agency (e.g. sharing it with others, using virtual agency or lending agency to natural phenomena).

This study uses a corpus stylistic approach to investigate the language of responsibility in pre-crime narratives written by mass murderers. The corpus consists of diaries and manifestos, and videos (video blogs) written in English by five high profile perpetrators (see table 1) in the months and years leading up to their massacres. All of the offenders being studied acted without the organisation of a wider terrorist group and with minimal help from others. The actions of such perpetrators are difficult to prevent or predict ahead of time because their profiles and behaviour before they commit crimes are not necessarily distinctive (Gill et al., 2014) and therefore any insight into their state of mind prior to their offences will add to research aimed at understanding and preventing such attacks.
Table 1. Corpus of Pre-Massacre Narratives

<table>
<thead>
<tr>
<th>Name</th>
<th>Event Description</th>
<th>Text Type</th>
<th>Word Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anders Breivik</td>
<td>Oslo bomb and Utoya massacre 2011</td>
<td>Manifesto</td>
<td>81,230 words</td>
</tr>
<tr>
<td>Eric Harris</td>
<td>Columbine shooting 1999</td>
<td>Diary</td>
<td>6647 words</td>
</tr>
<tr>
<td>Dylan Klebold</td>
<td>Columbine shooting 1999</td>
<td>Diary</td>
<td>5840 words</td>
</tr>
<tr>
<td>Elliot Rodger</td>
<td>West Coast US massacre of women 2014</td>
<td>Manifesto</td>
<td>107,958 words + 8 video diaries.</td>
</tr>
<tr>
<td>Sebastian Bosse</td>
<td>German school shooting 2006</td>
<td>Diary</td>
<td>1906 words</td>
</tr>
</tbody>
</table>

The corpus stylistic approach has traditionally been applied to fiction (e.g. Mahlberg, 2007) and uses corpus techniques to identify variations in style within or between texts. However, Hunt and Carter (2012) suggest that the same techniques can be applied to non-fiction texts and may elucidate the world view of a narrator by revealing patterns of language. In this study, after dividing the corpus for each perpetrator into sub-corpora of key events, variation in the language of the perpetrators is diachronically analysed by looking at key words, phrases and concepts, and their collocation networks in each sub-corpus. This allows us to explore the way in which each offender moves linguistically between different categories of agency over the years and months leading up to their crimes. These categories relate to constancy of the self over time, to what extent they consider themselves to be the same or different to others, and whether they present themselves as able to act on the world or controlled by it (Bamberg, 2011). The results demonstrate that changes in their sense of agency do not necessarily align with existing assumptions relating to crime and responsibility.

References


Designing, Compiling and Sampling Specialised Corpora for Researching Metaphor in Translation: The Case of SCoPE

Jesús Meiriño-Gómez (Universidade de Vigo, Spain)

This poster details the design criteria and compilation protocol, together with the procedures used in the sampling of excerpts, for the Specialised Corpus of Popular Economics (SCoPE). SCoPE includes two different sub-corpora, amounting to around 500,000 words each. One corpus is an English-Spanish parallel corpus, which consists of five American-English books and their Peninsular-Spanish versions. The other is a Spanish monolingual corpus, made up of seven Peninsular-Spanish works. The books covered by both collections address, to a greater or lesser extent, issues related to the economic shifts during the Global Systemic Crisis. They were written respectively by American and Spanish leading experts in the field, and published between 2010 and 2014. SCoPE was originally created within the framework of an on-going PhD research study, which aims to investigate how conceptual metaphors and their linguistic realisations are handled in translation, and describe the similarities and differences that might exist between translated and original texts in the target language.

A brief contextualisation to the research project is provided, followed by a detailed description of the different steps taken and challenges faced throughout the design, compilation and sampling process. The first stage of the project consisted of setting up the design criteria based on the purposes of the study, which allowed the definition of the target population and the sample frame. After specific books were selected for the corpus, they were converted into machine-readable format to allow the subsequent use of corpus analysis tools in their treatment and analysis. This conversion process yielded some changes to the texts in terms of format, content and structure, something which required a subsequent manual proofread. Given the goals of the study, certain elements such as editorial information, page numbers, figures and charts were removed as they were considered irrelevant. In addition and for the same reason, all of the texts were manually segmented at sentence level to make the subsequent consultation and analysis using AntPConc and AntConc easier, while the English and Spanish text pairs of the parallel corpus were also aligned at sentence level by hand. Due to the considerable size of the corpus and its sub-corpora, manual analysis of the complete texts was not possible. For this reason, excerpts taken from the original books were selected at random and then manually analysed, which allowed for a subsequent study by concordancing the previously obtained data. Details on the innovative methodology followed in the selection of excerpts, which involved statistical calculation methods, will be provided. The methodology described above, despite being framed in a particular research study, can be easily applied in other corpus-based studies of a similar nature.
Form and Function in Native and EFL Learners’ Academic Writing: What is the Difference?
Tsung-Ming Wu (National Taiwan Normal University, Taiwan)

An indicator of professional writing is the employment of lexical bundles, or a sequence of word combinations (Ådel & Erman, 2012). According to the definition, lexical bundles are recurrent word combinations and they “show a statistical tendency to co-occur” (Biber et al., 1999, p. 989). Previous research on lexical bundles has shown that native or professional writers tended to employ more lexical bundles than non-native or student writers (Ådel & Erman, 2012; Chen & Baker, 2010; Lin, 2011) and that their bundle use was different from non-native or student writers (Ådel & Erman, 2012; Chen & Baker, 2010; Cortes, 2004; Lin, 2011; Wei & Lei, 2011). However, earlier studies tended to adopt texts of different genres or of different disciplines for comparison, which may raise the issue of comparability. Another issue lies in the inconsistency of their comparing criteria; either their corpora size was inconsistent, with one corpus being much larger than the other, or their dispersion criterion differed among the corpora.

To address the comparability problem, the current study examined the employment of lexical bundles between the texts of the same academic genre, MA theses, by both native academic writers and the advanced EFL academic writers. The frequency threshold was set 40 times/ million words and the dispersion rate, 20% (i.e. 6 texts). This study attempted to explore the most frequently employed four-word bundles in the two corpora and to explore the form and function difference in the writing between the advanced EFL writers and native writers. It is assumed that if EFL academic writers are informed of a more professional use of lexical bundles, their writing skills are likely to improve.

A corpus of EFL MA theses and a corpus of native MA theses, which amounted to 1.3 million words, were therefore constructed for the current study. The EFL MA thesis corpus was comprised of 30 MA theses in the field of applied linguistics by Taiwanese writers. A total of 0.7 million words were included. The native MA thesis corpus, on the other hand, was made up of 30 MA theses in the field of applied linguistics. The native MA theses were collected from the graduate programs in the United States and the United Kingdom. To ensure the nativeness of the authors, their names and acknowledgement were referred to. The total number of words in this corpus was 0.6 million.

While previous studies have shown that native writers tended to employ more lexical bundles both in terms of the number and in terms of the type in their language production, the present study demonstrated the opposite pattern: EFL writers employed more lexical bundles both in number (EFL writers: native writers= 5,052: 1,576) and in type (EFL writers: native writers= 93: 42) than native writers as shown in Table 1. Overall, EFL writers demonstrated more uses of lexical bundles both in number and type. However, when the shared bundles between the two corpora were examined, EFL writers displayed an underuse trend...
(23 underused bundles out of the 29 shared bundles found in the two corpora).

Table 1 Corpus results between native speaker theses (NS theses) and non-native theses (NNS theses)

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Text</th>
<th>Corpus size</th>
<th>Freq. Criterion</th>
<th>Dispersion Criterion</th>
<th>Types before adjustment</th>
<th>Types after refinement</th>
<th>Tokens before refinement</th>
<th>Tokens after refinement</th>
</tr>
</thead>
<tbody>
<tr>
<td>NS theses</td>
<td>30</td>
<td>617,497</td>
<td>29</td>
<td>6</td>
<td>50</td>
<td>42</td>
<td>1,965</td>
<td>1,576</td>
</tr>
<tr>
<td>NNS theses</td>
<td>30</td>
<td>713,469</td>
<td>25</td>
<td>6</td>
<td>114</td>
<td>93</td>
<td>6,127</td>
<td>5,052</td>
</tr>
</tbody>
</table>

The present study also analyzed the structure and function difference between the two constructed corpora. Lin (2011)’s structural framework, a modification of Biber et al. (1999)’s structural framework, was adopted for the analysis of the bundle structure. Hyland (2008)’s functional framework was adopted for the analysis of the bundle function.

An examination of the bundle structures revealed that even though EFL writers employ more Noun Phrase (NP) bundles and Preposition Phrase (PP) bundles, which were often perceived as productive academic bundles, in terms of the number and the type, the percentage of NP and PP bundles in their writing was not as high as native writers’, as shown in Table 2. To exhibit the relative usage of the NP and PP bundles by native professional writers, a NS reference was compared. The NS reference was compiled by Lin (2011), which compiled 200 research articles from prestigious journals in applied linguistics.

When the function types of the bundles were investigated, EFL learners were found to employ lower percentage of the participant-oriented bundles despite the more varied bundle types and more bundles in number were found in the EFL learner corpus, as shown in Table 3.

An analysis of the structure type and function type of lexical bundles revealed that with different perspectives, the results differed. From the lens of percentage, native writers displayed similar bundle usage to professional writers, while in the raw frequency level, EFL writers seemed to have better command of bundle use, if more bundle uses represented higher language proficiency.

Table 2 Lexical bundles categorized according to structure types

<table>
<thead>
<tr>
<th>Category</th>
<th>NS</th>
<th>NNS</th>
<th>*NS reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>NP</td>
<td>14 (33.33%)</td>
<td>26 (27.96%)</td>
<td>33.77%</td>
</tr>
<tr>
<td>PP</td>
<td>17 (40.48%)</td>
<td>32 (34.41%)</td>
<td>43.05%</td>
</tr>
<tr>
<td>VP</td>
<td>6 (14.29%)</td>
<td>21 (20.43%)</td>
<td>20.53%</td>
</tr>
<tr>
<td>Others</td>
<td>5 (11.90%)</td>
<td>16 (17.20%)</td>
<td>2.65%</td>
</tr>
<tr>
<td>Total</td>
<td>42</td>
<td>93</td>
<td></td>
</tr>
</tbody>
</table>
Table 3 Lexical bundles categorized according to function types

<table>
<thead>
<tr>
<th>Function</th>
<th>NS</th>
<th>%</th>
<th>NNS</th>
<th>%</th>
<th>*NS reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Research-oriented bundles</td>
<td>17</td>
<td>39.53%</td>
<td>43</td>
<td>45.74%</td>
<td>37.70%</td>
</tr>
<tr>
<td>2. Text-oriented bundles</td>
<td>18</td>
<td>41.86%</td>
<td>41</td>
<td>43.62%</td>
<td>46.40%</td>
</tr>
<tr>
<td>3. Participant-oriented bundles</td>
<td>8</td>
<td>18.60%</td>
<td>10</td>
<td>10.64%</td>
<td>15.90%</td>
</tr>
</tbody>
</table>

An examination of the highly overused bundles (e.g., *on the other hand*) among the shared bundles illustrated some learner usage problems. Also, the examination of the highly underused bundles (e.g., *it is possible that, as well as the, and the fact that the*) suggested that teaching and learner style might play a role in the use of lexical bundles. The lexical bundles with significant difference were shown in Table 4. Based on the findings, some pedagogical implications are discussed for English learners and instructors.

Table 4 Lexical Bundles Employed with Significant Difference between NS and NNS

<table>
<thead>
<tr>
<th>Rank</th>
<th>Bundles</th>
<th>NNS Freq</th>
<th>NS Freq</th>
<th>LL</th>
<th>Sig value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>on the other hand + the</td>
<td>317</td>
<td>58</td>
<td>15.73</td>
<td>****</td>
</tr>
<tr>
<td>2</td>
<td>in the present study + the</td>
<td>302</td>
<td>68</td>
<td>6.31</td>
<td>*</td>
</tr>
<tr>
<td>3</td>
<td>the results of the</td>
<td>149</td>
<td>88</td>
<td>-21.05</td>
<td>****</td>
</tr>
<tr>
<td>4</td>
<td>at the end of + the</td>
<td>108</td>
<td>67</td>
<td>-18.22</td>
<td>****</td>
</tr>
<tr>
<td>5</td>
<td>at the same time</td>
<td>62</td>
<td>39</td>
<td>-10.98</td>
<td>***</td>
</tr>
<tr>
<td>6</td>
<td>at the beginning of + the</td>
<td>62</td>
<td>49</td>
<td>-22.08</td>
<td>****</td>
</tr>
<tr>
<td>7</td>
<td>the extent to which</td>
<td>56</td>
<td>28</td>
<td>-3.91</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>in the form of</td>
<td>55</td>
<td>45</td>
<td>-21.52</td>
<td>****</td>
</tr>
<tr>
<td>9</td>
<td>there was no significant difference</td>
<td>51</td>
<td>41</td>
<td>-19.03</td>
<td>***</td>
</tr>
<tr>
<td>10</td>
<td>the results of this</td>
<td>50</td>
<td>42</td>
<td>-20.97</td>
<td>****</td>
</tr>
<tr>
<td>11</td>
<td>as well as the</td>
<td>48</td>
<td>61</td>
<td>-51.76</td>
<td>****</td>
</tr>
<tr>
<td>12</td>
<td>on the basis of</td>
<td>46</td>
<td>28</td>
<td>-7.26</td>
<td>**</td>
</tr>
<tr>
<td>13</td>
<td>the fact that the</td>
<td>43</td>
<td>51</td>
<td>-40.23</td>
<td>****</td>
</tr>
<tr>
<td>14</td>
<td>in the field of</td>
<td>43</td>
<td>27</td>
<td>-7.57</td>
<td>**</td>
</tr>
<tr>
<td>15</td>
<td>i don’t know</td>
<td>43</td>
<td>27</td>
<td>-7.57</td>
<td>**</td>
</tr>
<tr>
<td>16</td>
<td>the meaning of the</td>
<td>42</td>
<td>30</td>
<td>-11.19</td>
<td>***</td>
</tr>
<tr>
<td>17</td>
<td>the purpose of the</td>
<td>41</td>
<td>38</td>
<td>-22.03</td>
<td>****</td>
</tr>
<tr>
<td>18</td>
<td>to the fact that</td>
<td>40</td>
<td>27</td>
<td>-8.94</td>
<td>**</td>
</tr>
<tr>
<td>19</td>
<td>it is possible that</td>
<td>39</td>
<td>52</td>
<td>-46.28</td>
<td>****</td>
</tr>
<tr>
<td>20</td>
<td>no significant difference</td>
<td>39</td>
<td>34</td>
<td>-18.00</td>
<td>****</td>
</tr>
<tr>
<td>21</td>
<td>significant difference between the</td>
<td>39</td>
<td>28</td>
<td>-10.55</td>
<td>**</td>
</tr>
<tr>
<td>22</td>
<td>in the case of</td>
<td>37</td>
<td>35</td>
<td>-20.88</td>
<td>****</td>
</tr>
<tr>
<td>23</td>
<td>can be seen in</td>
<td>34</td>
<td>27</td>
<td>-12.27</td>
<td>***</td>
</tr>
<tr>
<td>24</td>
<td>it should be noted</td>
<td>31</td>
<td>25</td>
<td>-11.67</td>
<td>***</td>
</tr>
<tr>
<td>25</td>
<td>there was a significant</td>
<td>30</td>
<td>28</td>
<td>-16.39</td>
<td>****</td>
</tr>
</tbody>
</table>

Note: LL = log likelihood; += NNS overuse; -= NS overuse; LL>3.84 → *p<.05; LL>6.63 → **p<.01; LL>10.83 → ***p<.001; LL>15.13 → ****p<.0001
References


Cortes, V. (2002). Lexical bundles in freshman composition. Using corpora to explore linguistic variation, 131-145.


Citing external sources in Educational Neuroscience articles: in search of an interdisciplinary stance and voice
Natalia Muguiro (University of Birmingham, Argentina)

Research on interdisciplinarity in academic discourse is growing in importance (Oakey & Russell, 2014; Petrić, 2006; Teich & Holtz, 2009; Thompson, 2015). Mostly, though, contrasting disciplines are compared with each other rather than with truly interdisciplinary writing. Starting from an understanding of writing as ‘dialogical’ (Bakhtin, 1981), the aim of this paper is to distinguish the ‘voices’ of external authors and the ‘stance’ taken towards their contributions (Thompson, 2012) in the interdisciplinary field of Educational Neuroscience through the study of citations in a corpus of research article introductions. So as to provide some parameters of comparison, two sub-corpora have been added to the interdisciplinary corpus. These consist of research article introductions from the two single-domain disciplines that give rise to the interdiscipline: Neuroscience, a ‘pure hard’ discipline, and Education, a ‘soft applied’ one (Becher & Trowler, 2001). Thus, this study might contribute to the description of the epistemic nature of interdisciplinary discourse.

One hundred and twenty research article introductions (152,202 words) from the three mentioned disciplinary fields have been analysed. Bibliographical citations and every other linguistic feature signalling the presence of external sources have been identified and analysed manually. Cases of ‘indirect attribution’ (Williams, 2010, p. 621), known as ‘attribution without citation’ (Thompson, 2005, p. 38), have been also considered and counted as individual citations. In addition, all citations that do not convey attribution but are part of averred statements have been counted too.

Thompson (2012) sees both ‘voice’ and ‘stance’ as ‘existing on a number of levels, ranging from the proposition through to the whole text’. For the purpose of this work, it is important to pay attention to the propositional level, through which we would ‘distinguish the voice that is responsible for a given statement’, being this the writer or ‘some other source’, and also ‘the stance that the writer takes towards a given proposition’ (Thompson, 2012, p. 119). A theoretical framework has been developed that distinguishes three parameters: the ‘strength’ and the ‘visibility’ of the external sources, and the ‘credit’ given to previous research. Once these individual parameters have been analysed, different patterns of parameter co-occurrence have been identified and described.

‘Strength’ refers to the strength of the external author’s voice. The strongest author’s voice will be represented by a citation that introduces an attributed proposition embedded in a -that clause. In contrast, the weakest author’s voice will be represented by a citation embedded in an averred statement. As for the ‘visibility’ of the external source, the distinction between ‘integral’ and ‘non-integral’ citations (Swales, 1990) must be considered first, before referring to the degree of ‘personalization’ or ‘objectification’ of the cited source (Hood, 2011; Coffin, 2009). The more depersonalised and/or objectified the sources, less visible they become.

‘Writer stance’ (Coffin, 2009) towards the attributed proposition is the position taken by the article writer ‘in relation to the words, observations or viewpoints’ presented by the referenced source. This is used to measure the amount of ‘credit’ given by the writer to the external author’s words, and, consequently, to the authors themselves. For instance, an ‘endorse’ stance gives more credit to the attributed proposition while a ‘contest’ stance gives less credit to the external author’s words or assumptions. The
amount of ‘credit’ given to external voices embedded in averred, rather than attributed, sources is not an easy feature to describe. In those cases, we might find citations used to describe a research process or procedure carried out by the external author or citations used to summarise results found or claims made by the external authors. In none of these cases is the external author responsible for the proposition; in fact, it is ‘the writer who is responsible’ (Thompson, 2005, p. 36) for the summary of the findings or the report of the process carried out. So, if we consider the amount of credit given to those averred sources, summarised findings and claims would be given more credit than described research processes or procedures, taking into account the distinction made by Hyland (2002) between higher and lower ‘communicative risk’ involvement on the part of the writer.

After the three parameters are described in each corpus, ‘citation patterns’ are identified and described. In the context of this study, I define a ‘citation pattern’ as a particular combination of the choices available from the three parameters described above. However, the description of this interweaving of choices is not an easy task, since we need to combine different categories that belong to different taxonomies, each of which highlights a different aspect. In fact, as Hunston (2011, p. 34) explains, taxonomies ‘tend to be complicated’ because they deal with several questions at the same time: not only with the question of ‘who is the speaker of the proposition’, but also with issues such as ‘how that speaker is represented’ and whether the author ‘delegates or reclaims responsibility for the proposition’. Different citation patterns can tell something about the authorial presence of the external sources in a text. For instance, we might encounter a citation pattern in which this presence is the most powerfully marked by the inclusion of a ‘human’, ‘identified’ source embedded in an ‘integral’ citation in which the writer takes an ‘endorse’ stance towards the attributed proposition, as in the following example:

(1) Indeed, Brankaer, Ghesquière, and De Smedt (2011) found that children with MID had impairments in numerical magnitude processing (Educational Neuroscience).

Clearly enough, the voice is the strongest (‘attribution’ through a verb + that clause) and the most visible (‘integral’, ‘human identified’ source) and the attributed proposition is given the most credit (‘endorse’ stance).

At the other extreme, a citation pattern in which a ‘non-integral’, completely depersonalised source (only present for ‘identification’ purposes) is part of an ‘averred’ statement, where no process is explained and no findings are summarised and the proposition is expressed in the words of the writer, would show the least powerfully marked authorial presence of the external voice, as in the example that follows:

(2) Cocaine facilitates the induction and maintenance of different forms of LTP in several areas of the brain including the hippocampus (Thompson et al., 2004; Del Olmo et al., 2006b; Huang et al., 2007a). LTP is a flexible event that can be erased by subsequent low frequency stimulation (LFS) in the hippocampus (Bashir and Collingridge, 1994) (Neuroscience).

Between these two extreme cases, there are many combinations that occur in different proportions and with different degrees of prominence. However, certain citation patterns are more or less frequent than others in the different corpora. This suggests important disciplinary and interdisciplinary differences in the ways in which the authorial voices of external sources are manifested in a text.
When articles from Neuroscience and Education are compared, preliminary findings show that Neuroscience sources are given more credit but are less visible, and that Education sources are more visible but are given less credit. As for strength, Education sources are slightly stronger. When citation patterns are considered, similar findings are observed. These results are predictable from what is already known about the difference between hard and social sciences: educators make other sources more visible but they give them less credit because they are entering into a debate with previous authors; neuroscientists, on the other hand, give more credit to other researchers’ findings but do not enter into a debate with them; rather, they ‘stand on their shoulders’.

When the findings obtained for the Educational Neuroscience corpus as a whole are compared with the Neuroscience and Education texts, several conclusions can be drawn. First, the external voices in Educational Neuroscience texts are slightly stronger than in Education and considerably stronger than in Neuroscience. Second, these voices are more visible than in Neuroscience but less visible than in Education. Third, and perhaps more importantly, new patterns that are not observed in the two pure disciplines emerge as typical features of the interdiscipline, such as longer blocks of attribution. These blocks connect pattern sequences whose possible general effect is to make external sources more powerfully marked. Finally, the discovery of new ‘citation patterns’ might lead to similar conclusions to those reached by other interdisciplinary studies which do not deal with citations. For example, Thompson (2015) concludes that interdisciplinary researchers write papers for a ‘broader audience composed of researchers from several disciplines’. Thus, the members of this new interdisciplinary audience are not all necessarily experts on the topics dealt with. The evidence found in the present study also highlights the presence of distinctive citation patterns that respond to this need to reach a broader audience of non-expert readers. These other external voices are part of the ‘imagined readership’ that appears to be connected not only with the world of academia but also with the world of work. In this way, the study of interdisciplinary audiences through the presence of external voices other than cited sources appears as a promising topic for future research. In a way, we would be giving rise to the study of new audiences for new disciplines.

References


Thompson, P. (2015). *Writing for an Interdisciplinary Audience: Corpus Perspectives* [Powerpoint slides]. Talk given at the *Federal University of Santa Catarina, Florianopolis, Brazil, 15th of April, 2016*.

1. Introduction

CoMParS is a resource under construction in the context of the long-term project German Grammar in European Comparison (GDE) at the IDS Mannheim. The principal goal of GDE is to create a novel contrastive grammar of German against the background of other European languages. Alongside German, which is the central focus, the core languages for comparison are English, French, Hungarian and Polish, representing different typological classes.

Unlike traditional contrastive grammars available for German, which usually cover language pairs and are based on formal grammatical categories, the new GDE grammar is developed in the spirit of functionalist typology. This implies that, instead of formal criteria, cognitively motivated functional domains in terms of Givón (1984) are used as _tertia comparationis_.

The purpose of CoMParS is to document the empirical basis of the theoretical assumptions of GDE-V and to illustrate the otherwise rather abstract content of grammar books by as many as possible naturally occurring and adequately presented multilingual examples, including information on their use in specific contexts and registers. These examples come from existing parallel corpora, and our presentation will focus on the legal aspects and consequences of this choice of language data.

2. Motivation and general assumptions

Corpus-based data are particularly precious because they present real context- and register-related language usage. At the same time, however, they are often not suitable for exemplification purposes because they are too complex and contain big portions of irrelevant material. The existing parallel corpora can already serve as a solid data source for contrastive research, but due to missing or sparse linguistic annotation, they do not reach their full potential. This is particularly true of multilingual parallel corpora: only two of them, namely InterCorp with 38 languages and 1.5 billion tokens (Čermák and Rosen 2012) and ParaSol with 31 languages and 27 million tokens (Waldenfels 2006) are lemmatized and grammatically annotated.

However, linguistic annotation in InterCorp and ParaSol identifies morphosyntactic properties alone; functional-semantic annotation is not available. Moreover, grammatical information available in these corpora is language-specific and thus not directly comparable across languages. As a result, only language-specific form-based queries can be performed, as opposed to meaning- or function-related queries such as, for example, “How is REFERENCE / QUANTIFICATION / REFLEXIVITY / POSSESSION / EXPERIENCE etc. expressed in languages L1...Lr”? However, precisely this kind of research question is substantial for crosslinguistic studies conducted within GDE. Obtaining parallel sequences appropriately exemplifying
specific communicative functions across different languages from the available corpora using form-based queries is a laborious procedure. Moreover, the extracted data are usually more complex than needed to illustrate the point at hand, and for this reason, they are not suited to be directly utilized as examples.

It is against this background that CoMParS is being constructed, as a small multilingual database of parallel sequences annotated with semantic-functional information and designed especially for the purposes of data-driven contrastive research, in particular contrastive grammar writing, with a view to language-didactic applications.

The general idea behind our approach to building CoMParS is to extract data out of the existing parallel corpora using the usual corpus query tools. The extracted data are then carefully examined by grammar writers, checked for quality of translation and accuracy of functional equivalence, as well as for relevance and applicability as examples in a (multilingual) contrastive grammar. We may call this a process of “refinement”.

Next, those parts of the selected aligned data are identified that are the smallest necessary for the appropriate exemplification purposes. Exactly those parts get additionally annotated with semantic-functional information. At the same time, it may happen that missing members of multilingual n-tuples are constructed (and clearly marked as such). This can be thought of as “enrichment”.

The extracted n-tuples also include information present in the original metadata and a link to the original resource from which they are cited.

3. Legal issues surrounding the construction of CoMParS

Intellectual Property Rights (IPR), in particular copyright and the sui generis database right have been identified as one of the major obstacles for the creation of language corpora. CoMParS is also heavily affected by these issues, as shown below.

Firstly, copyright needs to be addressed, because a large part of the material used in CoMParS is still under copyright protection (which expires 70 years after the death of the author). Indeed, in order to reproduce a copyright-protected work (including excerpts therefrom) and communicate it to the public, one needs an authorisation from the rightholder, unless the use enters within the scope of a statutory exception. CoMParS enters within the scope of the quotation exception under German law, and possibly other national laws of the EU Member States. International (art. 10.1 of the Berne Convention) and EU law (art. 5.3(d) of the Copyright Directive) require that in order to be lawful, quotation has to be justified by its purpose. In some jurisdictions, only excerpts of works can be quoted (e.g. in France), while others allow whole works to be quoted under specific circumstances (e.g. in Germany, where s. 51(1) UrhG allows ‘long scientific quotations’). Furthermore, most jurisdictions (including Germany and France) require that the citation is included in an independent work (i.e., that the quoting work meets the requirements for copyright protection, i.e. is its author’s own intellectual creation, even if the quotations were removed), which means that a mere compilation of citations, without any original contribution, is not allowed. While this would arguably exclude many language corpora from the scope of this exception, this is not the case of CoMParS, which, while based on data extracted from existing corpora, is also enriched/extended by additional annotations, functional alignments, and, in some
cases, also additional content, resulting in work derived from, and not merely copying, the original sources.

Moreover, the Court of Justice of the European Union recently ruled (C-145/10, Painer), that the Copyright Directive does not require the quoting work to meet the criteria for copyright protection. Member States are thus allowed to abandon this requirement (which is what Slovakia did in 2015), but are not obliged to do so. Even before the Painer case, courts in jurisdictions like France or Germany on occasions adopted an extensive interpretation of the quotation exception. This is what happened in the Germania 3 ruling in Germany (in which the Federal Constitutional Court ruled that copyright rules (and exceptions) should be interpreted in a way as not to inhibit basic freedoms, such as freedom of artistic expression or freedom of research) and the Microfor case in France (in which the Court of Cassation ruled that a work made for informational purposes — and (electronic) collections of linguistic data are likely to qualify in this category — can quote other works even if it does not itself meet the criteria for copyright protection).

Secondly, copyright may also protect compilations (such as language corpora) if the selection and arrangement of their contents is original (i.e. is its author’s own intellectual creation). In such cases, however, copyright only protects the ‘envelope’ (selection and arrangement), but not the contents. While CoMParS does reproduce data from other corpora, it does not copy their original selection and arrangement.

Finally, language corpora are likely to qualify as databases and therefore they may be protected by a *sui generis* database right if there was a substantial investment involved in their creation. The holder of the *sui generis* database right (i.e. the investor – in case of language corpora usually a research institution or a research funding agency) can prohibit extraction and re-utilisation of substantial parts of the database, regardless of its originality. While it is extremely unclear what constitutes a substantial part of a database, CoMPars only extracts and re-utilizes quantitatively small parts of the above mentioned corpora; therefore, the *sui generis* database right is not infringed.

### 4. Conclusion

The present contribution sketches the legal aspects of a decision to create an electronic language resource on the basis of quotations from other such resources, enriched by original content and structured according to semantic-functional criteria, in the context of EU law, and specifically German and French norms. We see this work as explaining the legal assumptions of CoMParS and, on a broader scale, as contributing to the search for legal solutions concerning the creation of complex language resources. We hope to have drawn the community’s attention to the potential use of the quotation exception (rather than merely the research exception) for the purpose of language resource construction.

### References

Legislative acts


Gesetz über Urheberrecht und verwandte Schutzrechte (German copyright act, UrhG).

Zákon č. 185/2015 Z. z. Autorský zákon (Slovak copyright act).

Court cases

Cour de cassation, Assemblé plénière, 30 October 1987, no. 86-11.918 (Microfor).

Bundesverfassungsgericht, 29 June 2000, 1 BvR 825/98 (Germania 3).

Court of Justice of the European Union, Third chamber, 1 December 2011, C-145/10 (Painer).
Aims and objectives

This paper describes the design, collection and evaluation of a comparable set of online corpora written in English, and influenced by four target languages: Arabic, Chinese, Persian, and Russian, in addition to a comparison corpus of native online English. The corpora have been designed to act as a resource for the automated identification of native language features, primarily for forensic purposes, but also with a view to being deployed for language learning uses. Ultimately, the research question this paper seeks to answer is as follows:

RQ: Can one successfully create a set of resources for online NLI which appropriately models target language users?

Background

The creation of resources for the task of native language identification (NLI) is by no means a new phenomenon. Both spoken and written corpora have previously been created for the purpose of second or foreign language teaching and learning and utilized for the purpose of NLI (Kochmar, 2011; Koppel et al., 2005; Tsur & Rappoport, 2007; Wong & Dras, 2009). A number of well-established resources that have been utilized as such include the International Corpus of Learner English (ICLE) (Granger, 2003), the Longman Corpus of Learner English (Pearson, 2016), the NUS Corpus of Learner English (NUCLE) (Dahlmeier, Ng & Wu, 2013), the Cambridge Learner Corpus (see Nicholls, 2003 for details), the McCall Corpus (Deutschmann et al., 2009), and the TOEFL11 corpus (Blanchard et al., 2013). Other studies have utilized data scraped from the web for NLI purposes (Brooke & Hirst, 2011). However, Hughes et al. (2006) cite the need for a standardised data set on which to base NLI evaluations.

The present paper seeks to address Hughes et al.'s (2006) recommendations and build on previous studies by moving away from the focus on constructing resources around learner errors, relying heavily on corpus contributors of University age, and obtaining data drawn from student essays, which is characteristic of the ESL and EFL resources used for NLI research. Further, the paper seeks to build on data scraping approaches by employing a sample frame design and a range of sources, which are lacking in such studies. In addition, unlike some previous NLI studies, the paper will allow for the elimination of features that also occur in native English by introducing a native English comparison corpus. Utilizing best practice from previous studies, the present paper incorporates the use of proficiency criteria, quality checking and expert judgements.
Materials

The paper is based on the collection and evaluation of three corpora:

Working corpora: four small corpora of English language blog posts, blog comments and forum comments influenced by the native languages of Arabic, Chinese, Persian, and Russian, respectively. These corpora were collected via a manual process by experts in the four respective native languages, and as such are small scale, but ‘pure’ (i.e. free from interference from other languages and accurately selected). The purpose of collecting the working corpora was to provide an accurate base from which to draw on for larger scale collection, and ultimately, for the automated extraction of features.

Validation corpora: four large corpora of English language blog posts, blog comments and forum comments influenced by the native languages of Arabic, Chinese, Persian, and Russian, respectively. These corpora were collected via a series of automated methods by corpus linguists, and as such are large scale, and in so being, introduce the potential for interference from other languages and inaccuracies in selection. One aim of the current paper is to evaluate the extent to which interference, inaccuracies and bias emerge.

Native corpus: a large reference corpus of native British English blog posts, blog comments and forum comments. This corpus was collected using the same automated methods as the validation corpora (or VC). The purpose of collecting the NC was to provide a suitable comparison (or control) corpus against which to measure the validation corpora.

Procedure

A proportional sampling frame was built for the corpora to appropriately model the target populations of British English, Arabic, Chinese, Persian and Russian internet users. The sampling frame was based on demographic data for the countries where the four validation target languages of Persian, Chinese, Russian and Arabic, and the native target language of British English are primarily spoken. Author-based demographic statistics were drawn from a range of demographic surveys, while text-based criteria of type (forum post, blog post, blog comment) and topic (drawn from the IPTC topic scheme) were equally weighted across the corpora.

The three forms of corpora outlined above were then collected. The working corpora were collected manually via Google searches for the exact phrases "my L1 is [TARGET LANGUAGE]", "my native language is [TARGET LANGUAGE]", "my mother tongue is [TARGET LANGUAGE]", "my mother language is [TARGET LANGUAGE]", or "my native tongue is [TARGET LANGUAGE]", coupled with the term "blog" or "forum". Data was collected by native language experts in each of the target validation languages of Arabic, Chinese, Persian and Russian. Each text was only selected for inclusion if it was judged by the language expert to be an authentic occurrence of an individual self-identifying as being a speaker of a given target language. The language experts collected data dating back 3-5 years as a maximum, in order to ensure that features contained in the data reflected current or recent
usage. The validation and native corpora were collected by corpus linguists, who formalised the working corpus collection criteria and sources, thus drawing from a standardised set of common blog and forum platforms, using a formalised set of native language and nationality identifying phrases and an automated approach.

Following collection, data were accessed for their quality with regard to i) data accuracy (the extent to which data reflected authentic instances of a given native language and was free from interference), and ii) goodness of fit (the extent to which the collected data matched the requirements of the sampling frame). 50 texts from each of the five corpora were selected at random (25 forum texts and 25 blog texts for each corpus) for this purpose.

Findings

The general collection methodology reveals that blog data in which authors self-identify as native speakers of a given language is more prolific than forum data of this nature. Productivity analyses suggest that certain blog (Wordpress and Blogspot) and forum platforms (Proboards) are more productive than others. Likewise, nationality identifier queries are far more productive than native language queries. Meanwhile, the quality analyses demonstrate that the resources are, for the most part, authentic and free from interference, though the same analyses do reveal the difficulties in matching online data to a proportional sampling frame due to the extent of missing metadata in online environments. The results suggest that the corpora will be more successful in aiding the identification of features associated with author gender than author age and education, since metadata is significantly lacking in the latter areas. With regard to text-based criteria such as topics, the Persian and British English samples demonstrate the greatest diversity of topics, while Arabic and Chinese demonstrate the least variety. Four of the five corpora are heavily weighted towards the topic of Lifestyle and Leisure (particularly Chinese), while Persian is weighted towards the topics of Social Issue and Technology. Collectively, the findings suggest that the resources show promise, but that further development work is required.

Conclusion

With appropriate sub-sampling to obtain a closer matching between collected data and the sample design, or targeted collection of additional data to address under-represented aspects of the sample design, the native language identification resources created here have numerous applications. For example, they could be used within the field of second language teaching and learning to help students perfect their English language skills when communicating online, or in day-to-day settings. Previous resources used for native language identification have tended to focus on formal or academic language used in essays, which only assists students with one aspect or environment in which they might use the English language.

A second application of such a resource is within the field of forensic linguistics, for example, in assisting with the task of author identification. Establishing features associated with differing forms of online English could help one to approximate the first language or nationality of an individual of unknown origin, while a study of features associated with differing aspects of the corpora's metadata
could further allow one to build a profile of an unknown author's gender, age, educational background and proficiency. The data also allows for the study of individual differences in online communication styles.

To collect further data, this formalised set of criteria could be used to obtain data going further back in time, in order to track changes in features over time. More generally, it is hoped that this methodology can be taken as a model and replicated for the collection of similar corpora on other languages, and that the study therefore goes some way to addressing Hughes et al.'s (2006) expressed need for a standardised set of criteria within NLI research.

References


NomVallex: Valency Patterns of Semantically Classified Czech Nouns
Veronika Kolářová, Jana Klímová and Anna Vernerová
(Charles University, Faculty of Mathematics and Physics, Czech Republic)

1 Introduction

Corpus-based models for describing the syntactic environments of individual lexical items play an important role in corpus lexicography, as demonstrated for example by the corpus-driven approach to the lexical grammar of English called Pattern grammar (Hunston & Francis, 2000) or the Corpus Pattern Analysis of Patrick Hanks (2013). While many valency lexicons are primarily intended for non-native speakers (e.g., Herbst et al. 2004), nouns are also covered in lexicons created mainly with NLP applications in mind, such as FrameNet1 and NomBank 1.02. Corpus-driven approach to valency of Czech nouns was applied by Čermáková (2009). Currently, valency patterns of Czech nouns are in focus of a new lexicographic project titled Corpus-based Valency Lexicon of Czech Nouns (Klímová, Kolářová, & Vernerová, 2016), using an acronym NomVallex3.

2 The development of NomVallex

NomVallex is a project building upon the theory of valency developed within Functional Generative Description (Sgall, Hajičová, & Panevová, 1986) and extending two existing valency lexicons developed within this tradition, Vallex (a valency lexicon of Czech verbs; Lopatková et al., 2015, 2016) and PDT-Vallex4 (containing valency patterns of verbs, nouns, adjectives and adverbs as they occurred in the Prague Dependency Treebank, Prague Czech-English Dependency Treebank and Prague Treebank of Spoken Czech; Hajič et al., 2003). Nouns to be included in NomVallex are selected based on the complexity of their valency patterns, special valency behaviour (e.g., special forms of participants, cf. Kolářová, 2014) and semantic class membership. Valency properties are captured in the form of valency frames for each meaning (lexical unit) of nouns included, and an enumeration of combinations of adnominal participants representing various valency patterns, as extracted from Czech corpora.

Currently, NomVallex follows the Vallex annotation scheme. Vallex was chosen as the base for the NomVallex project because it provides semantic class membership (Kettnerová, Lopatková, & Hrstková, 2008) and valency patterns for all meanings (i.e. lexical units)5 of verbs included, while PDT-Vallex covers only the lexical units that were encountered in the data of the treebanks in the Prague Dependency Family. Where possible, NomVallex maps nominal lexical units to their source verbal lexical units contained in Vallex. Technically, adopting Vallex software, NomVallex can serve as a supplement to Vallex, providing not only nominal entries but also links between the two parts of speech. Links to PDT-Vallex data are also recorded wherever they can be established.

---

1 https://framenet.icsi.berkeley.edu
2 http://nlp.cs.nyu.edu/meyers/NomBank.html
3 https://ufal.mff.cuni.cz/grants/nomvallex
4 http://hdl.handle.net/11858/00-097C-0000-0023-4338-F
5 Aspectual pairs (e.g., perfective odpovědět ‘to answer’ and imperfective odpovídat ‘to answer’) are usually treated within one lexeme (headword) in Vallex.
3 Extraction of valency patterns from Czech corpora

Czech is a highly inflectional language; valency participants of a word are primarily distinguished by their morphological category of case while the word order is very flexible, especially concerning valency participants expressed by prepositional phrases (PPs). Typically, an adnominal participant can be expressed by at least two forms (variants); in general, almost no combination of variants can be excluded. Therefore, searching for valency patterns of Czech nouns usually means searching for many various combinations of forms, including word order variants.

The following Czech lemmatized and morphologically annotated corpora are used: the synchronic part of the Czech National Corpus (CNC)\(^6\), the web corpus Araneum Bohemicum Maximum\(^7\) and the Prague Dependency Treebank (PDT 3.0)\(^8\). The PDT 3.0 is a small but manually syntactically annotated corpus, providing also semantic roles assigned to particular nodes (Mikulová et al., 2006). Using the CNC and the Araneum corpus, valency patterns of Czech nouns are being extracted either with the help of Sketch Engine’s Word Sketches (Kilgarriff & Tugwell, 2001), or by sophisticated CQL queries specified in the KonText application\(^9\). Searching through the PDT is carried out by the tool called PML-TQ (Štěpánek & Pajas, 2010).

4 Semantic classes in NomVallex and a preliminary list of entries

Nouns representing five semantic classes are included in NomVallex, namely Communication (e.g. odpověď ‘answer’), Exchange (e.g. dodávka ‘delivery’), Contact (e.g. dotyk ‘touch’), Mental action (e.g. dojem ‘impression’), and Psychological nouns (e.g. obava ‘fear’). The assignment of semantic class is carried over from Vallex: a noun is supposed to be assigned the same semantic class as its source verb in Vallex, with the exception of nouns that undergo a change in meaning. On the basis of the list of verbs in Vallex (see Table 1 for numbers of lexical units representing particular semantic classes), a preliminary list of noun entries was created. We aim to provide valency patterns of all types of Czech nouns with a meaning denoting an action or an abstract result of an action. These nouns are either derived from verbs by productive means (suffixes -(e)ní/tí, as in vykládání ‘explaining // unloading’ or pojetí ‘conception’) or by non-productive means including the zero suffix (such as vykládka ‘unloading’, výklad ‘explanation / interpretation’). The preliminary list of candidate entries to be included in NomVallex currently contains 1230 lemmas, cf. Table 2.

<table>
<thead>
<tr>
<th>Verbs in Vallex</th>
<th>Communication</th>
<th>Exchange</th>
<th>Contact</th>
<th>Mental action</th>
<th>Psychological verbs</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>428</td>
<td>182</td>
<td>125</td>
<td>338</td>
<td>143</td>
<td>1216</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Number of verbal lexical units in Vallex

---

\(^6\) http://korpus.cz/

\(^7\) http://ucts.uniba.sk/aranea_about/index.html

\(^8\) http://ufal.mff.cuni.cz/pdt3.0

As the first step, we carried out a quantitative analysis focusing on relative frequencies of combinations of participants modifying nouns representing the five selected semantic classes in the PDT 3.0. 623 such lemmas occurred in the PDT 3.0 in a total of 8273 occurrences (see Table 3).

The Graphs 1 and 2 show that the most frequent combination is the case when only Patient (PAT) is expressed (with the exception of non-productively derived nouns of Contact which represent the least frequent class and so the numbers may be influenced by their rare occurrence). The case when only Actor (ACT) is expressed is the second most frequent combination, followed by the combinations Actor + Patient or Patient + Addressee, the latter of which is applicable only in the case of nouns that have an Addressee (ADDR) in their valency frame. Interestingly, relative frequencies of the combination Actor + Patient are very low with nouns of Exchange and nouns of Contact. Relative frequencies of combinations of three participants – no more than 0,13% – are not shown in the Graphs.
Graph 1: Relative frequencies of selected combinations of participants modifying productively derived nouns in the PDT 3.0

Graph 2: Relative frequencies of selected combinations of participants modifying non-productively derived nouns in the PDT 3.0
6 Valency patterns of Czech nouns based on the CNC and the Araneum corpus

The current work focuses on the extraction of all possible forms of noun participants and their combinations from the CNC and the Araneum corpus, concentrating on nouns of Communication. The extracted corpus data give evidence about the following phenomena:

(i) Nouns derived from perfective verbs by productive means show slightly different valency behaviour than nouns derived from the corresponding imperfective verbs, even when both types of nouns denote an action.

(ii) There is a strong tendency for non-productively derived nouns for an increased number of possible expressions of their participants in comparison with their source verbs, cf. valency frames for the verb apelovat ‘to appeal’ and the noun apel ‘appeal’ illustrated in (1) and (2). Interestingly, despite the higher number, some of the forms are shared between particular participants, such as the prepositional phrase k ‘to’+DAT shared between ADDR and PAT in (3) and (4).

(1) apelovat  ‘to appeal’
   ACT(NOM;obl) ADDR( na ‘at’+ACC;obl) PAT(content_clause;obl)

(2) apel  ‘appeal’
   ACT(GEN, possessive;obl) ADDR(DAT, k ‘to’+DAT, na ‘at’+ACC;obl) PAT(k ‘to’+DAT, proti ‘against’+DAT, content_clause,inf;obl)

(3) apel k lidem.ADDR ‘an appeal to people’

(4) apel k ukončení násilí.PAT ‘an appeal to end the violence’

7 Conclusion

We believe our description of valency patterns of Czech deverbal nouns representing the five semantic classes will result in a valuable source of information, facilitating a detailed comparison of valency patterns of Czech nominal and verbal lexical units, providing also information about word-formation relations and semantics, including shifts in meaning.

Acknowledgements

The research reported in the paper was supported by the Czech Science Foundation under the project GA16-02196S. This work has been using language resources developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071).

References


In contrastive corpus-assisted studies the focus has recently shifted to the analysis of co-occurrence relations. “This new emphasis on the company words keep, to use Firth’s expression, has led to the discovery of a wide range of word combinations or multi-word units, which vary in fixedness and idiomaticity.” (Altenberg & Granger, 2002, p. 5; cf. also Ebeling & Oksefjell Ebeling, 2013, p. 50) We present a cross-linguistic study of semantic patterns whose function is to introduce a new phenomenon in the discourse situation, relating it to an already established location (cf. Johansson, 2007, pp. 234-5). Similarly to Hunston’s (2008) ‘semantic sequences’, the semantic patterns explored here are recurrent “sequences of words and phrases that may be very diverse in form and which are therefore more usefully characterised as sequences of meaning elements rather than as formal sequences” (Hunston, 2008, p. 271). The English semantic pattern to be dealt with can be described, in lexico-grammatical terms, as ‘location – presentation verb – new phenomenon’. Three syntactic types of presentative constructions have been described in literature quite thoroughly (e.g. Firbas, 1992; Ebeling, 2000; Johansson, 2007; Adam, 2013; Dušková, 2015):

i) full (existential) presentative constructions, e.g. There’s a long trip ahead of us;

ii) bare presentative constructions with a rhematic subject, e.g. a) A long trip is ahead of us, b) Ahead of us is a long trip;

iii) have-constructions (often with a locative subject, and possibly also with other verbs), e.g. a) We have a long trip ahead of us, b) The table has a lamp on it (Ebeling, 2000).

Depending on the grammatical structure, the new phenomenon is introduced on the scene either as the subject in pre- or post-verbal position (types ii and i, respectively) or as an object (iii).

Despite the differences in the syntactic structure of the English sentences, they can all be translated into Czech using the same construction: ‘locative adverbial – verb – new phenomenon’, i.e. Před námi je dlouhá cesta / Před sebou máme dlouhou cestu (lit. Ahead of us is a long trip / Ahead of us we-have a long trip), Na stole je lampa (lit. On the table is a lamp). In Czech, the topic–focus articulation is considered to be the primary factor determining the word order: given information tends to be presented at the beginning of a clause and new information at the end. Thus, the clause-initial Czech adverbial (typically expressed by a prepositional phrase, e.g. před námi, na stole) carries the information about the already established locative setting; the new phenomenon presented on the locative scene occurs in post-verbal position, functioning as the subject or object of the clause.

Our paper explores the possibility to employ the recurrent Czech presentative construction to track among its translation counterparts syntactically diverse English patterns which perform the same presentative function. Like Gast’s (2015, p. 5), our
method aims not only at describing the translation correspondences but also at making predictions about what syntactic constructions are likely to occur as implementations of the given semantic pattern.

We can assume that while languages have identical “needs of expression and communication... the means of expression vary from language to language” (Mathesius, 1936, p. 95; cf. also Haspelmath, 2010). The functional concept of ‘presentation’, shared by both languages, can therefore serve as a tertium comparationis. Since the Czech construction with a clause-initial locative element performs the function of presenting a new phenomenon on the scene, the question of how ‘presentation’ is expressed in English can be operationalised as ‘What English constructions serve as functional correspondences of the presentative Czech sentences (i.e. as their translation counterparts)?’ (cf. Gast, 2015).

Accordingly, we used the Czech clauses with clause-initial adverbials as potential markers or anchors (Malá, 2013; Gast, 2015) of the presentative function. This allowed us to identify among the English translation correspondences of the Czech presentative clauses various types of syntactic constructions whose function is to present a new phenomenon on the locative scene. We excluded the generally known ‘full’ and ‘bare’ presentative constructions (types i and ii above) from the study, and focused merely on the English constructions with a locative subject (iii b above).

A two-step procedure was adopted, relying on the Czech/English fiction sub-corpus of the bi-directional translation corpus InterCorp (2,710,000 tokens in the Czech section). In the first step, we proceeded from Czech sentences with initial locative adverbials (realised by prepositional phrases, e.g. na tváři in ex. 1) to their English translation counterparts. While in Czech there is a close link between the semantic role (location) and clause element function (adverbial), in English the looser ties between semantics and syntax lead to a more frequent employment of locative subjects. Such subjects serve as a context-bound scene on which new phenomena can be introduced (his face in ex. 1). Locative subjects occurred in four English syntactic patterns which can serve to present a new phenomenon on the scene. The SVO pattern (ex. 1) corresponds to the have-presentatives described above. The SVA pattern (ex. 2) is often mentioned as a result of the ‘locative alternation’ (Levin, 1993; Dowty, 2000; Fried, 2005) and our data demonstrate also its presentative function. Two additional, albeit infrequent, presentative patterns were attested in our data - SVpassA (ex. 3) and SVCs (ex. 4).

(1) Na tváři měl klidnej úsměv. (lit. On his face he-had peaceful smile) - His face wore a peaceful smile.
(2) V očích jí zářila klidná radost. (lit. In eyes to-her shined peaceful joy) - Her eyes glowed with peaceful joy.
(3) Na zdech byly zajímavé obrazy. (lit. On the walls were interesting pictures) - The walls were hung with interesting pictures.
(4) V jejich tvářích byla skutečná zloba. (lit. In their faces was real anger) - Their faces were full of real anger.

The presentative function is carried out primarily by the semantics of the ‘presentative’ verb. The wide range of English verbs which occurred in presentative constructions (exx. 1-4) can be classified into three classes:
a) ‘contain / be covered’ verbs (and copular predications), e.g. have, bear, contain, wear, hold, house; be bathed, coated, covered, decorated, dotted, filled, hung, inscribed, jammed, swathed, full, fragrant, rich, aglow (with/in/of) (exx. 1, 3, 4),
b) verbs of light, sound, and smell emission or movement, e.g. buzz, hum, ring, glow, flash, sparkle, reek, bristle, drip, swarm, twitch (ex. 2), and
c) ‘manifestation verbs’ related to sensory perception, e.g. show, display, radiate, read, say, e.g. The next page displayed a large wedding picture.

In the second step, the lexical verbs identified in step one were examined in English original texts in InterCorp (15,820,000 tokens). Among the various uses of the verbs, the ‘presentative semantic pattern’ represents “what is often said, not how... [the verb] is typically used” (Hunston, 2008, p. 291).

Our study demonstrated that the potential of the sentence to convey the presentation idea is primarily based on the presentation function of the verb. The verb also determines the syntactic structure that will be employed to implement the ‘presentation semantic pattern’. Apart from the presentation verb, the pattern always comprises the location (physical space, ex. 3, or body parts making it possible to add detail to the description of a person, exx. 1, 2, 4) and the bearer of a state or simple process (a specific entity, an abstract quality or its manifestation) introduced on the scene. In addition, in the SVA, SVPassA and SVCs patterns the interpretation of the locative element is often holistic, i.e. these patterns convey the sense that the location is fully affected by the state/action (exx. 2, 3 and 4).

From the methodological point of view, the study demonstrates how the use of a parallel translation corpus makes it possible to by-pass one of the limitations of corpus-assisted approaches in that it allows the researcher to search for patterns of meaning, or semantic patterns, rather than formal sequences.

References


**Sources**

1. Introduction

Modernism and social reform have been the subject of much debate in the predominantly conservative Saudi society. Within this debate, a number of ideological key terms have been contested between the two main groups: the conservatives and the progressives. One of the most significant of these is ‘allibraliah’, a loan word in Arabic that corresponds to ‘liberalism’ in English. In Laclau’s words, ‘allibraliah’ is considered an empty signifier, as different groups attempt to fill it with their own ideological meanings in order to gain or to maintain power (1996). In this talk, I will discuss how corpus methods can contribute to the study of the discursive construction and change of ‘allibraliah’ as an empty signifier. In addition, I will briefly show the means by which corpus analysis can be integrated with critical discourse analysis (Fairclough, 2001; Reisigl & Wodak, 2001) for an in-depth analysis of the mechanisms by which the construction of ‘allibraliah’ is changing.

The salient discursive analysis of the node ‘allibraliah’ follows Firth’s contextual theory of meaning, which views the meaning of a word as largely dependent on its relationship with co-occurring words (1957). For an empirical investigation of the usage of ‘allibraliah’, I employed Sinclair’s (1991) units of meaning model with a particular focus on analysing three types of lexico-semantic relations which are collocation, semantic preference and semantic prosody.

2. Data and methods

A corpus was compiled for this study, which consists of 575 articles (505,122 words) that contain the lemma ‘librali’ (liberal). These articles are taken from four Saudi newspapers published in the period between 2007 and 2016. Each article has been annotated for time, and part-of-speech tagged using Madamira tool for processing Modern Standard Arabic (MSA) texts developed by Pasha et al. (2014). The corpus was uploaded to the Sketch Engine corpus tool (Kilgarriff et al., 2004) in order to carry out a corpus analysis.

The designed methodological framework aims to answer the following research questions:

A) What are the fields of discourse associated with the discourse of ‘allibraliah’?
B) What are the semantic preferences around the term ‘allibraliah’?
C) What attitudes towards ‘allibraliah’ are held by Saudi newspapers and do they change over the whole period?

To answer these questions, an analysis of the word list was conducted to identify the top lexical items in the corpus and the meaning preferences associated with these words. This was followed by an analysis of the usage of these lexical items over time using trend analysis. Analysis of collocations was also conducted, to examine the words surrounding the term ‘allibraliah’. Finally, a concordance analysis of these
collocates was carried out for a close examination of the usage of the collocates with ‘allibraliah’ and the way they contribute to the construction of the meaning of ‘allibraliah’ over time.

3. Some findings

An initial analysis of the usage of ‘allibraliah’ over the period (2007-2016) has shown that it has been mostly used and contested in 2010. This finding emerged when reviewing the number of articles that constitute the corpus as seen in figures (1)

![Figure 1: Number of articles per year in the corpus (2007-2016)](image)

A close analysis of the discourses associated with ‘allibraliah’, performed by conducting a word list analysis has revealed that ‘allibraliah’ is associated with the discourse of religion (co-occurring with terms such as Islam, Muslim and Allah) and with the discourse of human rights (co-occurring with the use of terms ‘hurriah’ means freedom, ‘insan’/human and ‘haq’/right). The trend analysis for the usage of these terms over time has shown remarkable results in that words indexing the discourse of religion decrease in use, while terms indexing the discourse of human rights show an increase over the same period.

A look at the collocations of the term showed similar results, with the lexical items ‘Islam’ and ‘hurriah’ (freedom) being the most frequent collocates with ‘allibraliah’. The concordance analysis of these collocations over time reveals that ‘allibraliah’ was negatively represented at the beginning of the period, in which it has been described as a concept that contradicts Islam and that its association with freedom is a foreign negative concept. However, towards the end of the period, ‘allibraliah’ is positively represented, where its association with Islam indicates the possibility of its coexistence with the religion and where its association with freedom implies the possibility of creating a special version of liberalism for Saudi society.

Despite these informative results concerning the salient discursive construction of ‘allibraliah’, these findings still do not provide insights into the mechanisms by which these trends emerge within the social context. It is thus necessary to integrate the
corpus analysis with critical discourse analysis approaches. According to Baker et al. (2008), corpus analysis is not sufficient to explain why particular lexical patterns were found as it does not address the social, political and historical contexts. Therefore, critical discourse analysis of individual texts will be conducted to examine the way these changes are results of strategic use and the means by which these strategies are intertextualised in a chain of texts over time, and the relationship between these trends and the wider sociocultural context in Saudi Arabia over the period. This integration of the micro analysis will be illustrated with some examples.

4. Conclusion

Corpus analysis has revealed remarkable results about the salient features of the discursive construction of ‘allibraliah’ and the change of the discourse over time. It has shown the main discourses associated with ‘allibraliah’ and the shifts in these discourses. However, this informative corpus analysis of the discursive construction of ‘allibraliah’ would be more fruitful when integrated with an in-depth analysis of individual texts using critical discourse analysis approaches.

References


Multi-dimensional analysis (MDA) of register variation (Biber 1991; Biber & Conrad 2009) has proven its worth in the empirical study of English and a typologically varied handful of other languages (incl. Spanish, Chinese, Portuguese, Nukulaelae Tuvaluan, Korean, Somali). However, it has never been extensively applied to Slavic languages, which are known for their rich inflection, distinctive morphology (e.g. verbal aspect) and a fairly long literary tradition shaping the registers and styles of different genres and text types. This paper aims to discuss specific issues encountered when applying MDA to Czech (a West Slavic language), as well as to point to some methodological innovations we adopted while working on the project.

Although there is sufficient Czech language data (in terms of both extent and diversity) available for this type of research, no extensive investigation of this area has been carried out so far (with the exception of an unfinished and unpublished study by Vílem Kodýtek). Czech is also a highly interesting language for this type of research from the point of view of methodology: it exhibits a high degree of inflection (which goes hand in hand with abundant morphological variation) and a sociolinguistic situation bordering on diglossia (see e.g. Bermel 2014 for a recent overview).

Corpus compilation

The primary goal was to create a corpus as diverse and as representative of the wide range of uses of language as possible. At the topmost level, texts are classified into three modes of communication: written language, spoken language and internet communication. Each mode is further sub-divided into two or more divisions (e.g. the written mode has a fiction, non-fiction, journalism and private correspondence division), divisions then branch further into classes of texts. It needs to be emphasised that the classification mentioned above is not based on intratextual criteria (i.e. on the language used), but on extratextual properties of a text, namely its overall intended function (e.g. poem, scientific paper, column etc.).

The corpus consists of 45 classes of text, each represented by approx. 200,000 words, i.e. 9,074 mil. words in total (excluding punctuation). In order to achieve as diverse a composition of the corpus as possible, we decided to use text samples/chunks (instead of whole texts) of the same length (between 2,000–5,000 words). During the sampling procedure, we took into consideration that the beginnings, middle portions and endings of longer texts should be represented equally. The vast majority of texts in the corpus come from the Czech National Corpus's own resources (see http://wiki.korpus.cz/doku.php/en:cnk:uvod); genres/classes which were not covered by in-house language data production were acquired from other research centres focusing on collecting Czech linguistic data.

For almost all classes, we had more data than we needed (the exceptions being private correspondence and administrative texts). We sampled each class...
separately (stratified sampling), while paying particular attention to within-stratum diversity.

Features and their operationalization

Once morphologically annotated and lemmatized, the corpus was then searched for more than 140 features, ranging from phonology, morphology and word formation to syntax, lexicon and pragmatics. The list was partially drafted on the basis of an overview of the existing style and grammar literature on individual competing alternatives in Czech (e.g. Čmejrková & Hoffmannová 2011 or Čechová, Krčmová & Minářová 2008, to name but two), and partially on the basis of language variation findings resulting from designing applications for morphological annotation. Unlike other MDA studies, our list of features puts greater emphasis on morphological variation, lexicon-level variation, and type-based features (complementing the commonly used frequency-based characteristics).

The paper will focus on specific problems related to these features:

- frequency of noun cases – morphological features rely heavily on automatic morphosyntactic annotation, which exhibits a significant portion of false positives, especially in spoken and internet texts
- gender markers – the aim is to investigate the biases concerning texts by/about women by examining the variability between texts with different proportions of active female participants (N.B. grammatical gender in nouns is heavily lexicalized in Czech)
- type-based features (inventory of pronouns, prepositions, conjunctions) – as a complement to frequency characteristics of some POS categories, we also decided to include the size of the inventory (number of types) as a separate feature. The higher the number of different items used in a text, the more complex, situationally rooted and/or co-textually linked the text is. The use of uncommon grammatical words within a text of small size (up to 5,000 words) shows a high level of lexical richness.
- lexical classes (abstract nouns, taboo words, poetic words etc.) – problems pertaining to the definition of these classes and their precise delimitation (recall of the feature)
- lexical richness, thematic concentration, use of unigrams and bigrams – esp. the problem of normalization of the scores according to text/chunk length.

Preliminary results

The statistical evaluation of the features as measured on the individual texts is performed via exploratory factor analysis, which enables an interpretation of the multidimensional space (number of texts × number of features) using several two-dimensional scales (for an extended discussion of the methodology in the context of linguistics, see Biber 1991). Each of these can then be made sense of according to the features with which it correlates, e.g. as setting apart texts with a prevailing informative function from texts which are primarily subjective.

As of December 2015, the interim results show a clear separation between spoken and written texts, as expected. According to the features with highest loadings, the first dimension can be interpreted as differentiating dynamic/narrative
texts from texts where a descriptive approach prevails; the second dimension captures the difference between unprepared, situationally anchored texts and prepared texts unrelated to the situation of production/reception.

With regards to this top-level differentiation, some inferences can be ventured about the role and position of internet communication which are of utmost importance for the strategies of corpus compilation. When considering the first two dimensions, the web domain seems to overlap with the area of private correspondence (letters), fiction, journalism, non-fiction and (to some extent) also formal spoken communication. However, according to our preliminary results, informal spoken discourse does not seem to be replaceable by internet sources as far as recourse to specific linguistic means is concerned.

References

Bermel, N. (2014). ‘Czech Diglossia: Dismantling or Dissolution?’ In J. Árokay et al. (Eds.), Divided Languages? (pp. 21–37). Wien, Austria: Springer.


Effects of L1 Transfer on L2 Learners’ VN Collocational Use: A Corpus-based Study from Semantic Preference and Semantic Prosody Perspective

Lijuan Song (Zhejiang University of Technology, China) and Brent Wolter (Idaho State University, USA)

This study investigates whether L1 transfer occurs in L2 verb + noun (VN) collocational production at the semantic preference and semantic prosody levels. Semantic preference and prosody are two important indexes to show a verb’s collocational features in terms of collocate selection and the overall attitudinal tendency a verb and its collocates convey in context (Partington 2004; Paciorek and Williams 2015). Cross-linguistic comparisons were conducted to explore the different semantic preference and features between ten high frequency English verbs and their Chinese equivalents, followed by an evaluation of whether these cross-linguistic semantic differences have an effect on L2 learners’ VN collocational output. The data used in this study were collected from three corpora: the Corpus of Contemporary American English (COCA), the Beijing Language and Culture University Chinese Corpus (BCC), and the Ten-thousand English Compositions of Chinese Learners (TECCL) Corpus. The COCA and the BCC data were used to establish degree of overlap for semantic preference between translation equivalents of verbs in English and Chinese. Semantic prosody values were calculated for verbs in each language respectively. The TECCL was used to gather data on productive VN collocational output for the ten verbs for L1 Chinese learners of English as an L2. These productive collocations were then classified as either native-like or atypical though consultation of COCA data and the Oxford Collocations Dictionary in an effort to determine if the overlap values could be used to explain degree of native-like alignment (or misalignment) in learners’ productive use of VN collocations.

The results suggested that the semantic preference overlap between the English verbs and their Chinese equivalents had strong influence on L2 learners’ tendency to generate native-like collocations in English. Semantic prosody values, however, were much less reliable in predicting native-like collocations. A backward stepwise regression procedure revealed that the best fitting model was one that included only semantic preference overlap values as a predictor variable without either semantic prosody values $t(8) = 7.17, p < .001, R^2 = .87$. Further analysis showed a strong correlation between semantic overlap scores and native-like percentages $r = .93, p < .001$. These findings provide evidence for Jiang’s (2000) L2 lexical representation and development model, in which Jiang claims that L1 mediation takes place in L2 lexical production processes at a semantic and syntactic level. The present study indicates that L1 transfer occurs not only at the semantic and syntactic level, but also at the collocational level. In other words, the semantic preference features of a verb stored in one’s L1 mental lexicon were also activated and in effect in L2 learners’ VN collocational output process.

References


Using corpus methods to investigate teacher scaffolding in Special Education Needs (SEN) classrooms
Gillian Smith (Lancaster University, UK)

Children with developmental language disorders such as autism spectrum disorders (ASD) have difficulty interacting and expressing themselves coherently and efficiently. The individual’s communication skills affect their social and linguistic functioning, which has been positively correlated with academic attainment, psychiatric achievement and overall quality of life (Anderson et al., 2007; Garfin and Lord, 1986; NRC, 2001; Prizant et al., 2003). Hence, research has shown it is important to intervene from an early age in order to help children with communicative disorders develop the skills they need to thrive socially, academically and mentally (NRC, 2001). One primary form of treatment is language interventions, which provide opportunity for both the acquisition of knowledge and the development of communicative skills that allow social and linguistic progression (NRC, 2001). These interventions usually take place in the classroom, given that young children spend the majority of their time in school (Wilson, 2013). In Special Education Needs (SEN) classrooms, language interventions are vital to providing children with language and social skills, and much research has addressed SEN teaching methods. Such research, however, is often limited by its use of small-scale samples of classroom language data, which are then approached via manual analysis. This study addresses that limitation by applying a corpus-based method to the study of one teaching method (scaffolding) in SEN classrooms, thus enabling the exploitation of a larger and therefore more representative sample of language use.

The corpus created for this purpose consists of transcribed interactions between teachers and groups of between 3 and 9 pupils aged 11-16 years in a UK specialist school for children with SEN. All classroom activities recorded surrounded a shared reading or literacy activity led by the teacher. Whilst research on scaffolding and teacher-pupil interactions usually focuses on one-on-one interaction, in this case group interaction was deemed more relevant, because given the children’s language difficulties they would often opt not to participate in one-on-one interactions, making such interactions a less productive source of data. Shared reading activities were chosen because they provide an ideal scaffolding environment, one that allows teachers to provide a model whilst simultaneously allowing children to test their own responses and comprehension. Shared reading activities were recorded in two different classrooms in eight separate lessons, amounting to 8 hours and 52,813 words of spoken classroom discourse. A bespoke transcription system was utilised, based upon the schema devised by Gablasova et al. (2015) for the Trinity Learner Corpus. This is a contemporary set of transcription conventions that has been demonstrated as convenient for the construction of spoken learner corpora, providing a model under which both verbal and non-verbal details are transcribed, where meaningful.

Whilst Gablasova et al.’s (2015) model proved extremely useful, some adaptations had to be made to better fit my data. First, both Makaton sign language and the use of speaking communication aids were coded within my modified transcription scheme, as these were heavily used in the SEN classrooms and
represented a successful form of communication, which therefore must be included in transcription of classroom interactions. Second, the process for transcription of questions was reformed to incorporate an initial classification of question types (including non-interrogative clause questions) at the actual point of transcription, as a basis for subsequent analysis (see below). Finally, my modified transcription system encodes overlaps and interruptions, as these were also deemed meaningful in this data set.

The teaching method under analysis is scaffolding, which is among both the most widely used and the most commonly researched teaching methods. Scaffolding is a process involved in learning, in which supports from a more knowledgeable source allow a less knowledgeable participant to develop more complex cognitive skills than those they would be able to attain independently. In its simplest sense, as Reiser (2004:274) notes “the term scaffolding has traditionally been used to refer to the process by which a teacher or more knowledgeable peer assists a learner, altering the learning task so the learner can solve problems or accomplish tasks that would otherwise be out of reach”. To analyse scaffolding within the corpus, a necessary step is the definition and development of methods for searching for instances of features of teacher interaction that may indicate the use of scaffolding as a teaching method. Whilst there are multiple scaffolding features, I shall focus solely upon methods for identifying instances of the use of questions in scaffolding here, as this was the most prominent feature in the scaffolding literature (Palinscar and Brown, 1984, 1985; Puntambekar and Kolodner, 2005; Seymour and Helena, 2003; Winn, 1994).

One challenge is that in this literature, the features of scaffolding are often extremely ill-defined from a linguistic perspective. We are often given explanations of the functions of specific scaffolding techniques, but we see little description of their forms, which makes it difficult to search for them within a corpus. Thus, in order to define corpus queries upon which to build a suitable methodology, the first step is to move from these vague descriptions to grammatically sound definitions of linguistic forms. This process was informed by the grammars of Biber et al. (1999) and Huddleston and Pullum (2002). Using these grammars, working definitions for five question types were formulated: wh-questions, alternative questions, yes-no questions, non-interrogative clause questions and question tags. The grammars were also used to provide the linguistics structures of these question types. For example, a wh-question was defined as either:

\[
\text{wh-word} + V + S \text{ OR wh-word (S) } + V
\]

Following the definitions of these structures, the second step is to translate these grammatical definitions into corpus queries. In the case of the questions feature, this was done by translating the grammatical forms into a multiword regular expression query appropriate for CWB/CQPweb. For example, the linguistic structure of wh-questions was translated to the following CQP syntax:

\[
\text{[pos=":*Q:*" \& pos!="YQUE"] } [0,15] \text{ [word!="#" \& word!="%"]}[\text{word=":" } \& \text{ !"] within u}
\]
These queries were created through a trial and error process, and were successful to varying degrees. Some question types, such as wh-questions or question tags, were very easy to translate, whilst others, such as alternative questions, were considerably more complex and therefore problematic. This exercise underlines the non-straightforward nature of the process of query definition as a part of corpus methodology, and thus also the methodological contribution that this work can potentially make to the field of classroom language.

Finally, I apply the queries outlined to the SEN corpus to look at the use of questions in SEN classroom interactions. This is a two-step analysis: (i) looking at the frequency of different questions types across groups, and then (ii) also looking in more depth at the frequency of different elements within questions, such as types of wh-word. For example, the results on wh-questions include the following:

1. Wh-questions occurred in 32.8% of classroom utterances
2. 47.3% of questions were wh-questions
3. Teachers of lower ability children used wh-questions more than teachers of higher ability pupils
4. 63.4% of wh-questions began with determiners, 22.9% began with general adverbs, 12.6% with pronouns and 1.1% with degree adverbs

We can use these findings to make inferences about teacher-pupil interactions in SEN classrooms. For example, they suggest that lots of discussion revolves around concrete objects (hence the prevalence of determiner-based wh-questions) and that in particular this happens more when teachers address less advanced pupils. Finally, I undertake a comparison of the frequencies for various types of question found within the SEN corpus and those found by Biber et al. (1999) in general spoken English.

In sum, this research involves three key innovations. First, it introduces a novel application of corpus methodology in educational/developmental psycholinguistics. These results provide valuable new approach to analysing how scaffolding operates within the SEN classroom on a scale much greater than that of a handful of individual interactions. Second, it demonstrates how we may search this corpus for the features of teacher-pupil interaction (and specifically features of scaffolding) in an automatic (and therefore more objective/quantifiable) way. Third, the results of these searches allow us to look at the frequencies of different features of scaffolding (and different linguistic constructions of these features) within the data in order to make inferences about how teachers and pupils interact in SEN classrooms.

References


When people undergo traumatic events, they frequently turn to metaphor in an attempt to make what might initially seem indescribable into something comprehensible to others, and/or to help themselves reach a clearer understanding of what has happened to them. This investigation explores such metaphorical language produced in an online discussion forum by survivors of relationship abuse to communicate about various aspects of their experience.

The specific linguistic focus here consists of metaphorical analogies that such survivors use about various aspects of their abuse experience, including their perception of past actions, involved parties, present and future recovery, and emotions. The present analysis first explores the ways survivors ‘frame’ their experience through selection of a particular source domain, and then goes on to look at the various source domain ‘scenarios’ that are subsequently drawn upon to elaborate various salient details of the abuse. Further, the analysis discusses the ways in which survivors negotiate and develop metaphorical scenarios and frames among themselves in their forum discussions (see e.g. Musolff, 2016; Semino & Demjén, 2016 for discussion of frames and scenarios).

Metaphor is often used to discuss abstract, complex ideas in terms of more concrete entities, whereby certain real or perceived qualities from a concrete ‘source’ domain are mapped to an abstract ‘target’ domain. Cognitive linguists maintain that the metaphors we use in language offer evidence about how we conceive of the world around us. Systematic analysis of metaphors used in authentic discourse may thus provide particular insight into the values, attitudes, and experiences of relationship abuse survivors, especially as metaphor naturally lends itself as a resource when people are attempting to share, explain or make sense of highly emotional and/or traumatic events (see e.g. Demjén, 2016; Kövecses, 2000; Semino et al., 2015).

Primary material and methods

The empirical data for this investigation consists of the metaphorical language produced in a publically available online discussion forum for anonymous posters, accessible without password protection. The forum allows survivors of relationship abuse to start and/or respond to threads about their ongoing or past experience in an abusive relationship. The relationships discussed most frequently involve a love partner: a spouse, a live-in partner, or a boyfriend/girlfriend. Other abusive relationships discussed concern family members (usually parents, siblings, or children), colleagues, and/or platonic friends. The abuse under discussion is always mental and emotional, but is sometimes also manifested in other forms, e.g. physical, financial, etc. The forum threads comprise a corpus of 44.6 million words produced over a six-year period: 302,793 posts produced by 4561 individual posters in 4042 threads.
The corpus was first uploaded into Wmatrix, a web interface using the CLAWS part-of-speech tagger and USAS semantic tagger (see http://ucrel.lancs.ac.uk/wmatrix3.html). Concordance lines and co-text for all metaphorical comparisons overtly flagged by use of the lexeme ANALOGY were retrieved for analysis: roughly 500 extended metaphors in all. These comparisons were assigned brief ‘labels’ summarizing their contents, which were then semantically annotated to identify the underlying metaphorical frames focusing on “some aspects of a perceived reality [to] make them more salient in a communicating text” (Entman, 1993, p. 52). The particular scenario in any given frame was then fleshed out to fill in the correspondences between the various elements expressed in the analogy and the various elements in the abuse experience, along with any overall message. Finally, investigation into the co-text of the identified analogies allowed for evaluation into how survivors react to, accept, expand, and/or reject the metaphorical comparisons advanced by members of the discourse community.

By way of example, consider the metaphorical comparison in (1), labeled as a ‘tornado’ analogy.

(1) I think of them [abusers] kind of like tornados. They touch down and destroy everything in their path then disappear, we good people just happen to be the pretty red barn in the field they land in.

USAS annotation assigns the words ‘tornado’ with the code \textit{W4}, indicating that this term falls in to the discourse field of the \textit{world and environment} (\textit{W}), and more specifically, the \textit{weather} (\textit{4}). ‘Weather’ is then taken here as the frame. In this particular scenario, the abuser is equated with a tornado, while the survivors are simply in the wrong place at the wrong time, the random and disempowered victims of indomitable and ruthless forces of nature. Unlike many analogies that are discussed among posters, this particular contribution brought about no direct response.

\textbf{Sample analysis}

Three points immediately stand out upon analysis of the metaphorical comparisons in this corpus. First, there is immense variety in the types of selected frames, rather than any ‘one-size fits all’ solution. Second, few posters select the identical scenario to describe (parts of) their experience, even within the same frame. Third, despite such wide variety, most proposed comparisons strike a chord with other members of the discourse community; members frequently show their support for each other by complimenting posters for an analogy that particularly resonates. Only rarely is there dissent, although alternative analogies may be advanced; this is part of the negotiation between survivors as they try to reach an understanding of their experience.

Here we contrast a few selected scenarios belonging to the same frame, all explicitly marked at some point in the discussion as an analogy: twenty-eight analogies fall into the discourse field of the \textit{world and environment}, seven of which belong to the subdivision \textit{weather}. Within this frame, three survivors selected a scenario of natural disaster. Two elaborate on a tornado analogy, while the third discusses a hurricane analogy; see (1) above, (2), and (3) respectively.
In a blink of an eye we can loose all that we have - the psycho came through our lives much like a tornado showing no mercy on who they touched down on - but it is the love and volunteering of others much like this site and in our communities that pull victims together and get them back on their feet - How would we live without that love -we would be nothing but a tornado and a psychopath that only rips things apart.

My latest analogy for the abuser is, a hurricane came through and ripped the roof off my house - nothing I can do to change what happened, but I am in full control of my recovery.

The tornado analogy in (2) shares the same mappings as those already identified in (1). The remark in (2), however, is preceded by the poster's recounting of her then-current experience with an actual tornado that had just hit her local community. She describes the grief and shock among the people who had lost their homes and places of business, and – importantly – also the generosity and caring among those who were able to offer help and support. This image of solidarity in the face of inexplicable adversity adds an uplifting note to the analogy, missing in (1). The figurative hurricane in (3) plays the same disempowering role as the tornado in the two previous examples, yet here the poster describes her subsequent reclaiming of her own sense of empowerment in even stronger terms than in (2); the helpless victim of a natural disaster becomes a determined survivor.

This same contrast between disempowered victim and empowered survivor in the wake of a natural catastrophe is explicitly discussed by another poster, in an analogy about earthquakes – also falling into the ‘world and environment’ superordinate category, but annotated with the USAS tag for the subcategory of ‘geographical terms’ (W3 tag) rather than ‘weather’ (W4 tag). In (4), a poster brings up her therapist's analogy where the overall abuse experience is compared to an earthquake.

She gave me an analogy to state her position of total responsibility for choosing life's experiences. Her words, "If someone finds themselves in an earthquake, that person chose the experience on some level."

By the logic of this therapist, no abuse survivor is a completely innocent victim of random events, as is the contention in the earlier tornado and hurricane analogies. Instead, we are all at least partially responsible for everything that befalls us. In a sense, this therapist’s view is one of empowerment, for it should always be possible to actively change circumstances that are under our control. In this case, however, the poster adamantly rejects the implications of the therapist’s metaphor; see (5).

I can not accept this belief system. If in an earthquake, i do not hold myself accountable for being the victim of an earthquake, but only my actions following my experience. How i cope with it. i am not responsible for childhood abuse, only my journey to heal from it. And i did not choose to be a victim of a disordered abuser. But I am responsible for healing from it, growing from this so i will not be targeted again.
This statement thus mirrors the view expressed in the hurricane analogy in (3), with the distinction between hapless victim and responsible survivor. In these scenarios, while the posters accept no blame for the abuse itself, they do accept responsibility for their reactions to the abuse.

References


The Design and Development of *Corpas na Gaeilge Comhaimseartha* (Corpus of Contemporary Irish)

Katie Ní Loingsigh, Brian Ó Raghallaigh and Gearóid Ó Cleircín (Dublin City University, Ireland)

1 Introduction

*Corpas na Gaeilge Comhaimseartha* (Corpus of Contemporary Irish (CCI)) is a monolingual collection of Irish-language texts in digital format which was compiled in Fiontar & Scoil na Gaeilge, Dublin City University, and is available at the following link: [www.gaois.ie/g3m/en/](http://www.gaois.ie/g3m/en/). CCI currently contains 14.9 million words. Fiontar & Scoil na Gaeilge is a school in the Faculty of Humanities and Social Sciences in Dublin City University which specialises in interdisciplinary teaching and research through the medium of Irish and has developed the resources Téarma.ie, Logainm.ie, Ainm.ie and Dúchas.ie as well as other projects in language technology and digital humanities (Ó Raghallaigh & Měchura, 2014).

2 Background

There are five major corpora containing Irish-language data available for research and linguistic analysis at present. The primary Irish-language corpus currently available is *Nua-Chorpas na hÉireann* (The New Corpus for Ireland, (NCI)) which contains c.30 million words and has been annotated using a morphological analyser and a part-of-speech tagger as developed by Uí Dhonnchadha (2009). The core of this corpus is taken from an 8 million word corpus of Irish that was developed by *Institiúid Teangeolaíochta Éireann* (the Linguistic Institute of Ireland) as part of the EU PAROLE project (Kilgarriff, Rundell & Uí Dhonnchadha, 2006, p.133).

In addition to NCI, the Royal Irish Academy has compiled two Irish-language corpora, *Corpas na Gaeilge, 1600–1882* (Royal Irish Academy, 2004), which contains 7.2 million words and *Corpas na Gaeilge, 1882–1926* which contains 7.1 million words. These two historical corpora were compiled as part of lexicographic work undertaken for the Historical Dictionary of Irish project (Foclóir Stairiúil na Gaeilge, 2017). *Corpas na Gaeilge, 1882–1926* has been annotated and enriched using tools developed by Uí Dhonnchadha (2009) and Scannell (2009) and this work is still in progress (Uí Dhonnchadha et al. 2014). The fourth corpus, *Tobar na Gaedhilge* ([http://www.smo.uhi.ac.uk/~oduibhin/tobar/](http://www.smo.uhi.ac.uk/~oduibhin/tobar/)), is freely available and contains a collection of 20th century Gaelic texts, primarily in Irish, but also a limited number of texts in Scots-Gaelic. This corpus contains over 5 million words. A multilingual corpus of over 18 million words, *CELT – Corpus of Electronic Texts* ([http://www.ucc.ie/celt/](http://www.ucc.ie/celt/)), has been developed in University College Cork and includes texts in Irish, English, Latin and Anglo-Norman French and remains unannotated at present.

The Corpus of Contemporary Irish is the most recently developed Irish-language corpus and is currently unannotated. There are two main searches available to the user – a specific search ("This phrase as is") and a broad search, which provides approximate and related string matching. Results can be filtered according to collection.
3 Methodology and data collection

The primary aims of CCI is to provide a freely accessible linguistic resource which reflects current usage of the Irish language. The following design criteria were followed in the selection of material for inclusion in CCI:

- Eligible texts must be available digitally;
- Eligible texts must have been edited, i.e. material from blogs, social media etc. is ineligible;
- Prose texts (fiction and non-fiction) written and published in Irish from the year 2000 onwards are eligible;
- Translated material is ineligible.

Once texts are selected, the following data is cleaned before import:

- footnotes, glosses, long strings of text in other languages;
- bullet points and internal marks such as †, *, ◊, etc.;
- tables and figures;
- bibliographies (internal references in the body of the text are included).

At the outset, an initial list of Irish-language publishers and other copyright-holders was compiled and a brief outline of the project detailing its nature was given to each individual or organisation. Permission was sought to obtain edited texts in digital format and to use the electronic copies of the texts as part of the CCI project. As a result of these interactions, permission along with the necessary texts were received in various formats, i.e. primarily in PDF, Quark and Word format. CCI segments and related metadata are stored in UTF-8 encoding in a relational database and delivered through a web-based search interface. Although source CCI texts are stored on disk in XML format, they are not freely available to download in any standard XML format due to copyright restrictions. The collection and preparation of the various corpus sources will be examined in the following section of this paper.

The first major stage of development involved the ingestion of texts obtained from the published archive of Cois Life, one of the foremost Irish-language publishers, along with the ingestion of online material from Tuairisc.ie, Beo! and Nuacht RTÉ. This collection amounted to a preliminary corpus of 5.3 million words. The online material was automatically imported into CCI. The Cois Life archive was available in PDF format and was imported semi-automatically. This pilot version of CCI was made publically available in April 2016 while further material for inclusion in the corpus was being collected and formatted. The relevant metadata in relation to each of the collected texts were recorded in the corpus database as follows:

---

1 While every effort was made to stringently follow the above design criteria, there are still minor instances of strings of text in other languages, internal marks, etc. occurring in the corpus.

2 https://www.coislife.ie/

3 http://tuairisc.ie/

4 http://www.beo.ie/

5 http://www.rte.ie/news/nuacht/
The second major stage of development involved a significant expansion of the corpus in size and scope. New material was received from a number of significant copyright-holders and was manually analysed and prepared for inclusion in CCI. The current contents of the corpus are as follows:

<table>
<thead>
<tr>
<th>Type of text</th>
<th>Metadata recorded</th>
</tr>
</thead>
<tbody>
<tr>
<td>Page in published book</td>
<td>• Title;</td>
</tr>
<tr>
<td></td>
<td>• Author;</td>
</tr>
<tr>
<td></td>
<td>• Publisher;</td>
</tr>
<tr>
<td></td>
<td>• Year of publication;</td>
</tr>
<tr>
<td></td>
<td>• Page number.</td>
</tr>
<tr>
<td>Published articles from journals,</td>
<td>• Journal title;</td>
</tr>
<tr>
<td>newspapers, etc.</td>
<td>• Journal volume;</td>
</tr>
<tr>
<td></td>
<td>• Journal number;</td>
</tr>
<tr>
<td></td>
<td>• Year;</td>
</tr>
<tr>
<td></td>
<td>• Month;</td>
</tr>
<tr>
<td></td>
<td>• Article title;</td>
</tr>
<tr>
<td></td>
<td>• Article subtitle (where applicable);</td>
</tr>
<tr>
<td></td>
<td>• Author;</td>
</tr>
<tr>
<td></td>
<td>• Page number(s).</td>
</tr>
<tr>
<td>Online material</td>
<td>• Title;</td>
</tr>
<tr>
<td></td>
<td>• Subtitle (where applicable);</td>
</tr>
<tr>
<td></td>
<td>• Author;</td>
</tr>
<tr>
<td></td>
<td>• Publication date;</td>
</tr>
<tr>
<td></td>
<td>• URL.</td>
</tr>
</tbody>
</table>

Table 1: Metadata recorded

In conjunction with the collection and preparation of new material, the existing contents of the corpus were re-examined and analysed. A number of these texts in their entirety along with specific parts of other texts were judged to fall outside the design criteria and were cleaned from the corpus, e.g. Irish-language quotations from

---

6 Genre tagging of the texts in the corpus is currently ongoing. Once completed, this will provide a more detailed description of the various text types in CCI.
texts published during the 20th century or earlier, long strings of text in other languages, religious prayers, collections of proverbs, collections of poetry and grammatical texts among others. This manual re-examination served as a quality assessment of texts.

The manual conversion and preparation of texts proved to be the most labour-intensive aspect of the project. Additionally, the conversion of texts from various formats to .txt format created challenges as often the source text did not transfer exactly to .txt format. For example, the transfer of typesetting hyphens was a common issue in a significant number of texts converted to .txt format from PDF and Quark source files, e.g. tábh-acht, athbheo-chan, beag-nach. This occurred primarily in instances where restrictive columns in the source texts included hyphens to facilitate the design and layout of the text. These hyphens were manually deleted in these instances.

In addition to the inclusion of unnecessary hyphens, some non-printing control characters and also additional spaces between letters, words and paragraphs were transferred during the formatting of texts. Specific instances of letters being transferred incorrectly were also produced in certain texts, e.g. chúLghogarnail which is correctly spelled as chúlichogarnail. While every effort was made to ensure these minor inconsistencies did not transfer to the corpus, there are still, inevitably, a small number of examples of unnecessary hyphens, etc. present in CCI.

4 Conclusion and future research

This paper presents the initial phases in the compilation of the Corpus of Contemporary Irish. Apart from material still being processed, all of the Irish-language material received from publishers, individuals and organizations is now available online and further material will be added as it becomes available. It is intended to use CCI in conjunction with a corpus query tool to undertake linguistic research and analysis of grammatical and collocational patterns. With the aid of corpus query tools, the data will form a valuable resource for linguistic research on contemporary written Irish in areas such as stylistics, terminology, lexicography, phraseology, discourse analysis, etc.
References


Dimension of Twitter Trolling: Short Text Classification Using Multiple Correspondence Analysis
Isobelle Clarke (Aston University, UK)

Introduction

Despite the growing body of research on internet trolling (e.g. Donath, 1999; Herring, 2002; Shachaf & Hara, 2010; Hardaker, 2010; 2013; Whelan, 2013; Coles & West, 2016), still relatively little is known about the extent of its repertoires and linguistic properties. This paper reports on the findings of a project using Multiple Correspondence Analysis (MCA) to examine functional linguistic variation in 720 examples of Twitter trolling, with the aim to linguistically distinguish its different types.

Research on multi-dimensional text type analysis (MDA) shows that linguistic features will tend to co-occur in texts which are functionally and/or situationally similar (e.g. Biber, 1989). Typically, MDA takes the relative frequencies of many grammatical features from a number of texts in a language variety and subjects them to a factor analysis. The dimensions revealed from this are interpreted functionally and are then used to cluster the texts into distinct types. Due to the reliance on relative frequencies of grammatical features, most studies employing MDA have dealt with long texts (e.g. Biber, 1988) because it is difficult to accurately estimate the relative frequencies of grammatical features in short texts (Bijhold et al., 2010). Tweets are restricted to 140 characters in length (typically fewer than 30 words), which means that the relative frequencies of features are likely to be inaccurate. One way to deal with short texts is to concatenate them (e.g. Passonneau et al., 2014). However, this is not useful for identifying the functional linguistic variation between texts.

Instead, this study applies a new form of categorical MDA based on an MCA of the simple occurrence of a variety of lexical and grammatical forms in individual Tweets.

Methodology

‘Trolling’ is used in multiple contexts and describes numerous behaviours. This means that identifying examples of trolling is challenging, and largely depends on how the researcher defines it. Based on the understanding that words gain meaning through their use, I adopt Mihaylov and Nakov’s (2016: 403) definition of ‘trolling’: “those that have been called such by other people”, and use this to identify examples. Hence, if something is labelled as trolling, I take it to be such because each use of trolling contributes to its meaning. With this definition, the most inclusive approach for data collection would be to search for ‘trolling’ in Tweets, detect accusations and proceed to identify instances. However, this approach is labour-intensive and subjective as in multi-message and multi-user discussions, the researcher must decide what post was accused of trolling. To avoid this, I selected the imperative “stop trolling” as one of many possible search strings because, as a directive to stop the current behaviour, it is responsive, suggesting that Tweets prior to this instruction are instances of trolling. Using this search string, 720 Tweets were manually collected by extracting the posts which “stop trolling” was in reply to, or if the “stop trolling” Tweet quoted another
users Tweet, then the quoted Tweet was collected. For the latter, the quoted Tweet alongside “stop trolling” suggested that this quoted post was trolling. Following data collection, all of the Tweets were tagged using a Twitter Part-of-Speech (PoS) tagger developed by Gimpel et al. (2011).

Based on the tagged corpus, I then automatically identified occurrences of 86 features in the Tweets. These features are based on basic parts-of-speech, grammatical constructions (Biber, 1988), and additional features specific to trolling (Hardaker, 2013). This resulted in an 86 feature by 720 Tweet binary data matrix. Features occurring in less than 5 percent of the Tweets were removed, resulting in 62 linguistic features. Subsequently, MCA was performed on this data matrix in R using FactoMineR (Husson et al., 2017).

MCA is essentially a dimension reduction method, which aims to represent high dimensional categorical data into a low dimensional space. MCA is predominantly used to analyse questionnaire and survey data, however it has been used for linguistic purposes (e.g. Tummers et al., 2012; Glynn, 2009; 2014). In this study, each linguistic feature has two categories (i.e. presence and absence). The MCA assigns each category a positive or negative coordinate and a value indicating its contribution to the dimension (Le Roux & Rouanet, 2010). The MCA also returns a positive or negative coordinate to each Tweet on each dimension, which can be plotted to visualise the relationship between Tweets. Following Le Roux and Rouanet (2010), each dimension was interpreted by considering the variables whose contributions were above 0.81, the average contribution of a feature (100/124). Subsequently, Tweets with high positive and negative coordinates on each dimension were analysed to check and refine the functional interpretation. Finally, each Tweet’s dimension coordinates were plotted to examine if there were distinct types of trolling.

Results

The MCA was used to return three dimensions because they were readily interpretable and subsequent dimensions explained a limited amount of variance. The features most strongly contributing to these dimensions are presented in Table 1.

Dimension 1
Because the relative frequencies of features were not taken into account, the MCA may classify Tweets by length because typically more words means more features. To test this, each Tweet’s dimension coordinates were correlated to Tweet length. This revealed that Dimension 1 is strongly positively correlated to Tweet length ($r = 0.74$), Dimension 2 is moderately positively correlated ($r = 0.35$) and Dimension 3 is weakly positively correlated ($r = 0.19$). A closer examination of the linguistic features strongly contributing to Dimension 1 (Table 1) supported this interpretation as positive coordinates were assigned to the presence of features, whilst negative coordinates were assigned to the absence of features. For this reason, Dimension 1 is excluded from further analysis.

Dimension 2
The linguistic features strongly contributing to negative Dimension 2 have an interactive function. Second person pronouns suggest that the writer is interacting with a specific person. Question marks are associated with interaction as they indicate
a question is being asked, and *interjections* are inherently interactive because they are immediate responses to stimuli.

Alternatively, those features strongly contributing to positive Dimension 2 are associated with informationally dense Tweets. *Prepositions, numerals, quantifiers, proper nouns* and *attributive adjectives* are used to provide specific detail. *Nominalisations* are indicative of a high informational load (Biber, 1988). The fact that Dimension 2 has a moderately positive correlation to Tweet length supports this interpretation because informational Tweets tend to be greater in length.

Dimension 2 can therefore be seen to reflect Biber’s (1988: 107) “Informational versus Involved Production” dimension. This is supported with examples strongly associated to this dimension.

**Table 1:** The features strongly contributing to the Dimensions.

<table>
<thead>
<tr>
<th>Dim</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>WH-pronouns, Nominalisations, Prepositions, Past tense, <em>be</em> as a main verb, Other adverbs, Public verbs, Determiners, Amplifiers, Auxiliary <em>be</em>, Coordinating conjunctions, Quantifiers, Second person pronouns, Analytic negation, Third person pronouns, Other pronouns, Infinitives, First person pronouns, Subject Pronouns, WH-words, Possessive pronouns, Prediction modals, Contrastive conjunctions, Auxiliary <em>do</em>, It, Verbs of perception, Object pronouns, Private verbs, Accusative case, Conditionals.</td>
</tr>
</tbody>
</table>

- Absence of Nouns, absence of Prepositions, absence of Subject pronouns, absence of Other pronouns, absence of Past tense, absence of First person pronouns, absence of Second person pronouns, absence of Determiners, absence of Other adverbs, absence of Accusative case, absence of Private verbs, absence of Nominalisations.

| 2   | Nouns, absence of Accusative case, Prepositions, Attributive adjectives, Determiners, Past tense, absence of Second person pronouns, Articles, Predicative adjectives, *be* as a main verb, Proper nouns, absence of Other pronouns, Quantifiers, absence of Mentioning, Capitalisation, Nominalisation, Comparatives, Hashtags, Numerals, Quoting, Superlatives, Cause subordinators. |

- Absence of Nouns, Accusative case, Question marks, Second person pronouns, absence of Prepositions, Other pronouns, absence of Attributive Adjectives, Interjections, absence of Proper nouns, absence of Articles, Mentioning, absence of Past tense, Nominative case, absence of *be* as a main verb, absence of Nominalisations.

| 3   | WH-pronouns, absence of Predicative adjectives, absence of First person pronouns, Other pronouns, Past tense, absence of *be* as a main verb, absence of Subject pronouns, Second person pronouns, Hashtags, Nominalisations, Public verbs, Auxiliary *be*, Numerals, Time subordinators, Place adverbials, Sasive verbs, Verb-initial, Perfect tense, Question marks. |

- Predicative adjectives, absence of Nouns, Comparatives, *be* as a main verb, Subject pronouns, First person pronouns, Auxiliary *do*, Contrastive conjunctions, absence of Prepositions, Analytic negation, absence of Past tense, absence of Second person pronouns, Nominative case, absence of Question marks, absence of Other pronouns, absence of Nominalisation.
Dimension 3
The features strongly contributing to negative Dimension 3 have an attitudinal function. *Be* as a main verb, *predicative adjectives* and *comparatives* function to express attitudes towards specific things. The co-occurrence of *first person pronouns* suggests that a personal opinion is conveyed. This interpretation is supported by Examples 1 and 2, which are Tweets strongly associated to negative Dimension 3. Both Tweets express a personal opinion.

(1) @username I don’t know letoya but she’s better than Beyonce
(2) Rogue One was so bad I don’t know if I ever want to watch movies again

By contrast, the features strongly contributing to positive Dimension 3 have an antagonistic function. *Second person pronouns* suggests that the Tweets are targeted. The co-occurrence of *initial verbs*, *question marks*, and *public verbs* indicate that someone’s speech is being brought into question. Additionally, *place adverbials*, *time subordinators* and *nominalisations* suggests that there is a high degree of specificity. This interpretation is supported with Examples 3 and 4, which are Tweets strongly associated to positive Dimension 3. These Tweets contain questions antagonistically directed to users.

(3) @username Speaking of which, y’all are sleeping w/ Russia so much you’ve got bed sores. When’s he resigning?
(4) Probably from the hundreds of thousands of businesses and entertainers that produce promo shirts? Are you suggesting someone invented that?

In summary, this dimension represents Tweets functioning to express attitudinal judgement versus antagonistic Tweets. This dimension is in line with Hardaker’s (2013) and Merritt’s (2012) descriptions of trolling behaviours, specifically that trolls post provocative content and are inherently hostile.

Using R, each text’s Dimension 2 and Dimension 3 coordinates were plotted on a map to show where each Tweet lies in relation to these dimensions (see Figure 1). Figure 1 shows that there are no clear clusters, but rather there is a continuous range of linguistic variation.

Conclusion

Based on this analysis, two dimensions of linguistic variation have been identified: namely, *interactive* versus *informational*, and *attitudinal* versus *antagonistic*. The former dimension echoes Biber’s (1988) dimension, whilst the latter reflects Hardaker’s (2013) and Merritt’s (2012) definitions that trolling can be provocative and hostile. By plotting each Tweet’s Dimension coordinates, it is possible to see the continuous range of linguistic variation of trolling.

Due to manual extraction, only 720 examples were collected, which is, by today’s standards, a comparatively small corpus. Therefore, future research will work on automating this process with other efficient search strings so that more dimensions of linguistic variation can be revealed.
**Figure 1:** Plot of each Tweet’s Dimension 2 and Dimension 3 coordinates.

![Plot of each Tweet’s Dimension 2 and Dimension 3 coordinates.](image)

**References**


Coles, B. A. & West, M. (2016). "Trolling the trolls: Online forum users constructions of the nature and properties of trolling" *Computers in Human Behavior* 60: 233-244


Hardaker, C. (2013). “‘Uh....not to be nitpicky,..., but...the past tense of drag is dragged, not drug.’ An overview of trolling strategies” Journal of Language Aggression and Conflict 1(1): 58-86


Stylistics has long been focused on the excellence of technique, with the goal of exploring the relation between language structure and its function (Leech and Short, 2007). The literary approach of stylistics explains the choice of language in terms of its artistic effects, treating stylistics as a bridge between linguistic analysis to literary criticism (Carter and Simpson, 1989). On the other hand, the fast development of natural language processing and corpus-based statistical technology has made it possible to study linguistic patterns of literary texts in large quantities. As a result, this development furthers the study of some original concerns of authorial style, such as if an author’s style is as distinguishable as those critics have believed (Busse & McIntyre, 2010). Researchers have employed corpus and computational methods in the study of authorship attribution, however, most of these studies have focused on attributing unknown texts to possible candidates while rarely making attempts to identify the stylistic features contributing to the author’s distinctiveness. Thus, Hugh Craig (1999) argued that authorship attribution should go hand in hand with stylistic description. Our study draws on both literary stylistic description and the techniques used in authorship attribution to explore the stylistic distinctiveness of Edgar Allan Poe to construct the stylistic profile of a gothic writer with both stylistic descriptiveness and statistical rigor.

Being a world renowned figure of the 19 century American literature, Edgar Allan Poe left the world with a large quantity of literary works including essays, short stories and literary criticism that have received constant academic interest from different fields of study. Scholars from these fields vary in their research interests in terms of both language and literature. Some studies devote to interpreting the theme of a particular writing of Poe (Gargano, 1960; Ginsberg, 1998), while some aim at elaborating Poe’s literary and aesthetic theories (Polonsky, 2002; Freeman, 2013). Studies exploring the stylistic features of Poe’s writings (Carringer, 1974; Fisher, 2002), have often focused on the gothic elements of his stories and generally are from an impressionist qualitative approach. Doubt exists pertaining to whether critics’ artistic intuition can be supported by more objective evidence which remains salient in all Poe’s writings rather than individual sentences, and thus, the quantitative research of Poe’s writing is also needed. However, comparatively fewer studies have been carried out from this approach.

Using a computational stylistic approach via cluster analysis, our study explored the stylistic features of Poe’s short stories. This computational stylistic approach evaluated distinctive feature qualities and identified specific stylistic properties that distinguished Poe’s stories from other gothic writers’ work. Seventeen tales from seven world-renowned gothic writers---- Ann Radcliffe, Matthew Lewis,
Charles Brockden Brown, Edgar Allan Poe, Nathaniel Hawthorne, Robert Louis Stevenson, and Bram Stoker were selected for automatic cluster analysis. These works represent the gothic fiction in different eras of development from 1790 to 1909, with Poe's writings positioned in the middle of the timescale. Four consecutive experiments were conducted, in which, a variety of stylistic features were chosen as criteria for cluster analysis. Based on the chosen features, Poe's writings were successfully clustered together as separate from those of the other gothic writers, suggesting that the chosen features from Poe's writings can be distinguished from the other gothic writers. Following this, further comparisons between Poe's text collection and that of other writers were carried out to further explore Poe's characteristic stylistic features.

These results show that the distinctiveness of Poe’s style of writing could be captured by quantitative computational methods via cluster analysis. This distinctiveness manifested itself in a variety of stylistic features including bag of words, core lexical words, the part of speech of core lexical words, and semantic fields. To be more specific, Poe’s texts revealed a strong preference towards the preposition of “upon” over “on”, which contrasts sharply with the writings of the other gothic writers. In terms of the lexical words, Poe favors those denoting places and constructions, such as “wall(s)”, “vault”, “dungeon” and “floor”, which echoes with Carringer’s (1974: 508) remarks that Poe’s writings “conspicuously involve severely restrictive enclosures”. Moreover, the words denoting “anatomy and physiology” and “health and disease” are distinctively salient in Poe’s writings. The former category involves a great variety of body parts ranging from top (“head”, “face”, “neck” and so on) to bottom (such as “abdomen”, “foot”, “heel”) and from outside (such as “hair”, “limbs”, “wrists”) to inside (“lungs”, “nerves”, “blood” and so on), while the latter contains both the words of physical symptoms (such as “inflamed”, “pallor”, “fever”) and mental symptoms (such as “exhausted”, “agonies”, “stupor”), which help create a sense of psychological terror manifested in Poe’s gothic works. In terms of the part of speech of the core lexical words, Poe’s texts demonstrate a noticeable preference for general adverbs, which, with their modifying function, contribute to the expressiveness of Poe writings.

References


The present study starts from the observation that traditional lexicography has tended to rely on corpora of written text. It is hypothesized that this might be to the detriment of covering the commonest colloquial lexical units which carry the main burden of everyday conversation, are usually acquired early on in life and are therefore more deeply anchored in speakers' brains than units first encountered in the course of education.

This hypothesis receives support from a detailed examination of the treatment accorded four English and four French high-frequency words and three English and three French high-frequency phrasemes in ten different dictionaries as well as from a spot check on twenty medium-frequency phrasemes.

The methodology used for this purpose proceeded in seven steps. The first step involved generating several frequency lists, viz. a) a frequency list of the 3000 most common lemmas in the spoken portion of the *Corpus de référence du français contemporain* and a large corpus of spoken British and American English and b) three lists of the 3000 most common 3-, 4- and 5-grams; word frequencies and multi-word strings were identified using the relevant functions of the *Sketchengine*. In the second step, lemmas either labelled 'colloquial' in several dictionaries or identifiable as significantly more frequent in speech than in writing were extracted from frequency list A, and lexical bundles that did not constitute phrasemes were eliminated from frequency list B. In the third step, a random selection was made of four items from the first list and three items from the second list. In the fourth step, the frequency data obtained for individual lemmas were compared with native-speaker usage ratings. The fifth step involved a detailed analysis of the lexico-grammatical and pragmatic features of the selected items, following the corpus-driven approach to habitual co-occurrences of words ('usuelle Wortverbindungen') developed at the *Institut für Deutsche Sprache* (Steyer, 2009, 2013). This approach is based on three methodological premises which draw inspiration from the British tradition of text analysis established by Firth and Sinclair, viz. a) it derives structure from the data during the analysis rather than in advance; b) it foregrounds language as use; and c) it lets the data speak for itself, allowing the observer to form an unbiased picture of authentic language in use. The data are listed in terms of node words and their primary and secondary collocates (e.g. *never occurred to me* in the case of the English node *idea*) and are then subjected to thorough scrutiny with a view to determining the internal structure and typical variation found with node-collocate pairs and establishing the presence or otherwise of node-collocate pairs with similar characteristics. This fifth step thus involves two sub-stages (cf. also Hanks, 2013, p. 92): the first involves grouping the evidence into recurrent semantic-pragmatic patterns; the second is the assignment of meaning to each pattern.

The general finding is that current lexicographic descriptions of spoken French are often patchy and inadequate with respect to various lexico-grammatical features, while the description of spoken English is somewhat more advanced though far from satisfactory.
Most notably, there was found to be a dearth of information on the collocational range of colloquial items in most of the French and the English dictionaries under investigation, although, at least theoretically, native-speaker lexicographers could have retrieved some collocations from memory. There is on the whole a considerable uniformity in the content of both monolingual and bilingual dictionaries, with French monolingual learners’ dictionaries and bilingual dictionaries tending to adopt collocations and sense divisions from standard reference works such as the Petit Robert.

Since the extraction of collocations is an essential prerequisite for the determination of meanings, it is hardly surprising to find that the marking of sub-senses may not be sufficiently clear for the encoding needs of non-native speakers of either English or French. Here the clearest examples are the highly polysemous verb lâcher and the complex preposition autour de in French. The only way to enable learners to gain an overview of, and ultimately to make productive use of, a verb like lâcher is to opt for splitting up the various senses derived from its literal meaning ‘ne plus tenir’ rather than lumping them all together, as PR does. It would be unrealistic to expect learners to derive such specific uses as the following from the general sense, especially since this use features a non-human subject:

(1) C’est un film qui vous prend à la gorge dès le départ et ça ne vous lâche plus.

Another main finding concerns the rudimentary treatment of common multi-word items with a clear discoursal function. Thus, nine of out of ten dictionaries fail to record the sense in which the common French preposition autour de is used to indicate that someone or something is at the centre of a particular endeavour, and all the dictionaries give low priority to the use of French n’importe quoi as a discourse marker. Most seriously perhaps, there is no indication of the typical contextual embedding or common lexical collocations of discourse markers (e.g. - N’importe quoi. – Ah si.). The aforementioned spot check on twenty medium-frequency items shows that coverage differs between dictionaries and that differences between American and British English often go unrecorded.

Most of the examples found in the dictionaries under investigation illustrate written usage, and of those that illustrate spoken usage many lack some naturalness feature or other. There is some evidence that different types of exemplification may be needed for different words (cf. Hausmann, 2005). lâcher is one example of a low-collocability word which cannot be illustrated by means of typical co-occurrences. With such words, users will need a large number of (at least) sentence-length examples to grasp the various meanings of the word. The case is different with the French noun look, a clear example of a high-collocability word which requires little exemplification beyond information on collocation.

Three dictionaries deserve special comment: DAFLES, Harraps and Longman Dictionary of Contemporary English. Although it has several compensating strengths which fall outside the scope of this study (cf. Verlinde, Binon & Selva, 2006), DAFLES has almost no entries for colloquial words, a fact which may be due to its corpus base.

By contrast, Harraps achieves a remarkable harmonization of descriptive and pedagogic needs. A measure of the overall quality of this monument to bilingual lexicography is the inclusion of a large number of colloquial senses of both lâcher
and péter and the provision of illustrative sentences from which students may confidently extrapolate personal choices. Like all current dictionaries, however, even Harraps is still weak on discoursal items. Longman stands out from the other monolingual dictionaries in recording a fairly large number of colloquial collocations and in achieving broad coverage of multi-word markers.

There are two important theoretical lessons to be drawn thence. The first, which concerns corpus linguistics, is that medium-sized spoken corpora like the CRFC or the BESpoken corpus will shed light on lexical patterns and collocations about which even very large mega-corpora of written language are completely uninformative. This means that there may well be a second corpus revolution ahead which will apply Sinclair’s famous dictum that ‘the language looks rather different when you look at a lot of it at once’ to the investigation of intimate and colloquial language use.

The second theoretical lesson is that colloquial words, far from being stylistically ‘inferior’ substitutes of more formal words, are imbued with their own specific shades of meaning, phraseology, and pragmatics. It is as if there is a primary lexis which is even more deeply submerged in the routines of everyday life and thus even less accessible to native-speaker intuition than the secondary written lexis, but there can be no doubt that such lexis is communicatively prior and that its detailed description is of crucial importance to second or foreign language learners. The present study suggests that much of the primary lexis of French and English, and very probably other languages as well, remains almost undescibed in respect of many of its features, with dire consequences for foreign learners aspiring to acquire native-like proficiency. Reliance on the rarer and clumsier words or lexical units may make their language use sound stilted and unnative-like (e.g. il n’abandonne jamais son portable rather than il ne lâche jamais son portable).

References


PONS. *Großwörterbuch Französisch. Deutsch-Französisch/Französisch-Deutsch*. Stuttgart: PONS.


Sketchengine Retrieved from http://www.sketchengine.co.uk/. Lexical Computing CZ


From ICE to ICC: A proposal for an International Comparable Corpus

John Kirk (Dresden University of Technology, Germany)
and Anna Čermáková (Charles University Prague, Czech Republic)

There is broad agreement that the International Corpus of English project has been highly successful because it has facilitated numerous comparisons of L1 and L2 national varieties of English worldwide. Those comparisons encompass the lexical and morpho-syntactic structural levels, as well as comparisons of discourse types and written registers (cf. e.g. Greenbaum 1996; Hundt & Gut 2012; Aarts et al. 2013; and the papers in the Special Issues of World Englishes vol. 15(1) (1996) and vol. 36(3) (2017), to mention but a few key studies). No small part of this success rests with the fact that for each national variety there has been chosen a set of spoken and written text categories which are deemed to be representative of each national variety: 15 discourse situations (totalling 60%) and 17 written registers (totalling 40%). A major review of the ICE project has been undertaken and its results and outcomes are to be agreed upon at ICAME in Prague in May 2017. It seems likely that the text categories will be expanded to include electronic texts and some flexibility in text category choice will become possible.

At the same time, spoken and/or written corpora have been compiled for other languages (cf. list of non-English corpora in e.g. O’Keeffe et al. (2007: 294-296) or the non-English corpora discussed in Xiao (2008) or Ostler (2008)). Xiao makes comparisons with corpora of English: for instance, the Polish National Corpus replicates the structure of the British National Corpus (Xiao 2008: 387), as does the Czech National Corpus (Čermák 1997), which contains spoken texts similar to those of demographically sampled component of BNC (Xiao 2008: 388-389, Čermák 2009). However, no corpus of another language appears to be composed with the range and balance of text categories and quantities of texts as contained within an ICE corpus. The existing corpora in various languages are generally compiled on very different principles and do not allow direct cross-linguistic contrastive comparisons.

Corpus-based contrastive studies are a growing research area and researchers have voiced need for more rigorous analytical framework (e.g. Aijmer et al. 1996, Altenberg & Granger 2002, Marzo et al. 2012, Aijmer & Altenberg 2013, Altenberg & Aijmer 2013). The majority of contrastive studies are being carried out on two languages only, one of the reasons being the lack of comparable data. Contrastive analysis relies on two types of data (Granger 2003): translation (parallel) corpora and comparable corpora (cf. McEnery & Xiao 2007). While translation corpora contain original texts and their translations, comparable corpora contain original texts in two or more languages that have been selected on comparable criteria for text categories and quantities for each category, such as the Lancaster Corpus of Mandarin Chinese, which uses the same sampling frame of the Lancaster/Oslo-Bergen Corpus, or the Aarhus Corpus of Contract Law (both cited in McEnery & Hardie 2012: 19; cf. also e.g. Sharoff et al. 2014). Comparable corpora are an essential data source to support contrastive analyses, since the translation corpora are usually limited as far as text types are concerned (Johansson 2007).
What we are introducing is not a parallel translation corpus such as the *English-Swedish Parallel Corpus*, the *English-Norwegian Parallel Corpus* (ENPC), or the *InterCorp* corpus; rather, it is the creation of an International Comparable Corpus (ICC – pronounced to rhyme with *lick*) with as many languages as wish to come on board. Phase I will start with national, standard(ised) European languages. An expression of interest to collaborate on this project has been expressed for the following languages: German, French, Czech, Slovak, Polish, Finnish, Norwegian, Swedish, and Scottish Gaelic. The first collaborative meeting is to be held in June 2017 in Prague.

The ultimate goal of this project is the facilitation of contrastive studies between English and other languages involving highly comparable datasets of spoken, written and probably electronic registers. A striking and unique feature of each new corpus will be its substantial spoken component, at present comprising 600,000 words (or 60% of the current total). The revised ICE format, to be adopted here, is likely to safeguard this large amount of spoken texts but will include electronic texts as well. Such provision of spoken data across 15 or so discourse situations for contrastive analysis will be unprecedented and invaluable for future research. This will then also allow the much-needed cross-linguistic comparisons of spoken language, further investigations may include the area of pragmatics, such as pragmatic discourse markers (cf. e.g. Aijmer & Vandenberg 2006).

The proposed comparable corpus ICC will allow substantially to add to existing contrastive corpus-based research (e.g. studies of English-German contrasts, such as König & Gast (2012), or English-Norwegian contrasts, such as Ebeling & Ebeling (2013)), and will allow replicability and comparisons with other languages, i.e. a corpus-based empirical approach to each pair of contrasts, with spin-offs for the others, would all become possible. A further application will almost certainly be possible in bilingual lexicography (as shown by the papers in Sharoff et al. 2013).

Following the launch of *ICE Phase II* at the ICAME conference in Prague in May 2017 and the first ICC meeting in June 2017, *Corpus Linguistics 2017* seems an ideal and opportune moment to introduce this exciting new international, multi-lingual corpus project and to present at the outset some of the issues and challenges it raises as well as the solutions being adopted.

**References**


A bidirectional English-Portuguese Dictionary of Verbal Collocations
Stella E. O. Tagnin (University of São Paulo, Brazil)

It is by now widely accepted that phraseological units, mainly collocations, are essential for a person's fluency (Pawley & Syder, 1983; Lewis, 1993). By extension, their knowledge is fundamental to translators who should have an excellent command of the pair of languages they work with. However, there are hardly any bilingual reference sources where students or translators can search for unknown collocations in the target language, one exception being Benson & Benson's A Russian-English Dictionary of Verbal Collocations (1993). A survey (Tagnin, 1998) of how nine English verbal collocations and their Portuguese equivalents have been treated in nine standard dictionaries, both monolingual English and bilingual English-Portuguese (Sinclair, 1987; Procter, 1995; Longman, 1995; Vallandro & Vallandro, 1976; Houaiss & Avery, 1964; Houaiss, 1982; Macmillan, 2002; Jelin, 2006) (Oxford, 2002), and in three Brazilian dictionaries, two bilingual and one monolingual (Houaiss & Avery, 1964; Taylor, 1982; Ferreira, 1975) showed that these collocations have not received systematic treatment, which makes it very difficult to find the desired collocation.

Table 1: Verbal collocations looked up in various dictionaries

<table>
<thead>
<tr>
<th>English verbal collocations</th>
<th>Portuguese verbal collocations</th>
</tr>
</thead>
<tbody>
<tr>
<td>make/take a decision</td>
<td>tomar uma decisão</td>
</tr>
<tr>
<td>meet/satisfy a need</td>
<td>satisfazer uma necessidade</td>
</tr>
<tr>
<td>make one's living</td>
<td>ganhar a vida</td>
</tr>
<tr>
<td>put a curse on sb</td>
<td>rogar praga em alg.</td>
</tr>
<tr>
<td>dispel/resolve a doubt</td>
<td>dirimir uma dúvida</td>
</tr>
<tr>
<td>deliver a speech</td>
<td>proferir um discurso</td>
</tr>
<tr>
<td>take an exam</td>
<td>fazer um exame</td>
</tr>
<tr>
<td>do battle</td>
<td>travar (uma) batalha</td>
</tr>
<tr>
<td>wage war</td>
<td>travar (uma) guerra</td>
</tr>
</tbody>
</table>

First of all, they may be listed under the verb, which is most often the case, or under the noun. It might be appropriate to list a verbal collocation under the verb in a dictionary aimed at the comprehension of language but not in one aimed at production. Usually a speaker or writer knows the noun, the referential lexeme, but might not know the verb that goes with it. For that reason, verbal collocations should come under the noun, as Hausmann (1985) has already claimed, for the noun is the base in a verbal collocation. Besides, only very rarely are verbal collocations listed as an entry in their own right. Mostly, they are either listed as a subentry, or in the definition or still as part of an example, but sometimes not highlighted in any form.
This paper will discuss the compilation of a bidirectional English-
Portuguese/Portuguese-English dictionary of verbal collocations, which
ecompass the following structures:

\[
\begin{align*}
V \text{ (Det) } & N \quad \text{make trouble, make an impression} \\
V + \text{Prep} + N \quad & \text{come into force, keep in touch} \\
V + \text{Adj} \quad & \text{get rich, go wrong}
\end{align*}
\]

It will report on the early pre-corpus stages of the project, when examples were
collected from books, magazines and newspapers to the present time when
citations were extracted from corpora, mainly the *Contemporary Corpus of
American English* (COCA) (Davies, 2016), the *Corpus do Português* (Davies &
Ferreira, 2016) and the *Corpus Brasileiro* (Berber Sardinha, 2016). For some
specific collocations the internet was searched. The paper will also discuss
decisions that had to be made, such as (1) 'is this a collocation?'; (2) 'is
frequency the only criterion for a combination to be considered a collocation?';
(3) 'is this collocation too specialized to be included in a general dictionary?';
(4) 'is this a good example, does it make the meaning clear?'; (5) 'how can we
find a good equivalent'; (6) 'should cognate verbal collocations be included?';
(7) 'how do we account for verbal collocations for which the equivalent is not a
collocation?'

These are some of the answers we came up with:

1. For a combination to be considered a collocation it must present
some type of lexical restriction: *give a book* is not a collocation because *give*
in this sense means to hand something over to someone and it can combine with
any object, such as *give a book, give a pencil, give a present, give a dress* etc.
However *give a paper* at a conference has a very specific meaning and is thus
considered a collocation.

2. Frequency is not always the sole criterion, exclusivity (Brezina,
McEnery, & Wattam, 2015, p. 140) is equally relevant. The Portuguese word for
*doubt, dúvida*, collocates with ‘esclarecer’ (clarify) 2730 times in the *Corpus do
Português: Web/Dialect* but only 321 with its synonym *dirimir*. Nevertheless,
DÚVIDA is the most common collocate of DIRIMIR, which qualifies the
combination as a collocation. The same may be said for FURL which, though
not a frequent verb, combines mostly with some type of sail: *Scouts untied lines, furled sails, dropped anchor. I went aloft to furl the mainsail in a
blow.*

3. This can be a tricky question, but we have decided to only include
specialized collocations which are known to the general public or, as the editors
of the *Oxford Collocations dictionary for students of English* have put it, to the
instance, *score a goal* has been included, while *jump offside* has not. By the
same token, everyday legal collocations such as *file/settle/dismiss a lawsuit* are
also listed.

4. We have attempted to use examples that make the meaning clear
because the dictionary does not include definitions. For instance, *Tom, you
bring up a point that I brought up a few weeks back* does not offer enough
context for the user to infer the meaning of bring up a point, whereas While I doubt this is true, it does bring up a point I want to discuss does.

(5) Finding a good equivalent can be difficult at times for various reasons: the collocation in language A is not translated by a similar collocation in language B. Whereas crash a party is a V + N collocation, its Portuguese equivalent, ‘entrar de penetra numa festa’ is not actually a collocation as we have defined it. Some collocations are translated by a single verb, like go sour whose equivalent is simply ‘azedar’. More often than not nouns are quite different across languages: make arrangements becomes ‘tomar providências’ in Portuguese. In such cases one has to rely on one’s own knowledge of both languages or resort to searches in monolingual or even bilingual parallel corpora (Tagnin, 2007).

(6) We have opted to include cognate verbal collocations because they may differ in their inclusion or not of a determiner, for instance. Make a difference requires some kind of determiner, while in the Portuguese translation, ‘fazer (uma) diferença’, the determiner is not compulsory.

(7) Because this is a bidirectional dictionary, the verbal collocation will only be listed in the source language, that is, if it is a verbal collocation in English but not in Portuguese, there will be an entry for it in the English-Portuguese direction, but not in the reverse direction. For example, go sour will have an entry in the English-Portuguese direction with its equivalent ‘azedar’, but ‘azedar’ will not be an entry in the other direction as it is not a verbal collocation.

Here are two sample entries, one in each direction. Please note that the example for the equivalent in the target language is not a translation of the example in the source language; it is always an authentic rendition in its own right.

**English-Portuguese**

- advance [progress]
  - **advance, make an**
    - Residents have reported seeing Mr. Taylor’s forces, which had made a swift **advance** into the city centre on Wednesday, now withdraw from the area. 
    - ... and that in the essential things she had **made no advance**. British Lung Foundation research is **making dramatic advances**...
  - progresso, fazer
    - ... A China **fez progressos** significativos quanto a direitos humanos ...
  - avanços, fazer
    - **O jogador, além de marcar, também deverá fazer avanços periódicos pela ponta-direita. Precisamos fazer avanços significativos no sentido de torná-los mais fáceis de usar.**

- **advances, make ~ to**
  - I’d be glad if you’d stop **making advances** to my daughter. Her husband heard about the **advances** Simon **had made** to his wife.
  - cantada, passar uma
As mulheres ... em geral acabam percebendo que não precisam de homens nem para lhes passar uma cantada.

- advances, rebuff/ reject /resist sb.'s ~
...Jacobs, the slave who lived in the 1830s, described being so desperate to resist the unwanted sexual advances of her white enslaver, that...
- investidas, barrar/repudiar as; resistir às
Apesar de ter se mantido virgem ... resistindo firmemente às investidas de vários pretendentes... ... para barrar as investidas de um concorrente mais ou menos anônimo ...

Portuguese-English

avanço
- avanços, fazer
O jogador, além de marcar, também deverá fazer avanços periódicos pela ponta-direita. Precisamos fazer avanços significativos no sentido de torná-los mais fáceis de usar.
- advance, make an
Residents have reported seeing Mr. Taylor’s forces, which had made a swift advance into the city centre on Wednesday, now withdraw from the area. ... and that in the essential things she had made no advance. British Lung Foundation research is making dramatic advances...

References

Grappling with Shakespeare's words: maximizing historical corpus-based approaches
Jonathan Culpeper and Amelia Joulain-Jay
(Lancaster University, UK)

This paper reports some of the work being undertaken in the context of the Encyclopaedia of Shakespeare's Language Project (see Shakespearelang, n.d.). Funded by the Arts & Humanities Research Council (AHRC), this project brings the corpus approach into the heart of Shakespearean studies and more generally Early Modern English. It affords fresh insights into Shakespeare's use of language at multiple levels – words, phrases, semantic themes, character profiles and more. In particular, it reveals what Shakespeare's language meant to the Elizabethans through the analysis of millions of words written by his contemporaries. The main output of the project will be a two-volume Encyclopaedia, published in paper and online. This paper focuses on Volume 1, which essentially is a corpus-based dictionary of Shakespeare's words. We elaborate on (a) the problems we encountered and solutions we adopted, and (b) how our corpus-based approach improves on current Shakespearean language scholarship.

Working on historical data brings with it familiar problems for the corpus researcher. We will briefly elaborate on the following problematic areas and explain how we tackled them:

(1) Spelling variation. Spelling variation is not, of course, unique to historical data. In the present-day world of global electronically mediated communication, authors are hardly conforming to one spelling standard. Nevertheless, spelling variation is particularly marked in Shakespeare's period, partly because standardised spelling was only just beginning to emerge, but also because of printing practices (e.g. line justification was largely achieved by adding in extra consonants or the letter <e>, or using the wider <y> instead of <i>). Our solution was to use VARD (VARiant De tector) (Baron, n.d.), software which can help identify and standardise historical spelling variation. But all is not plain sailing. One key issue is: what do you regularize to? This is not a problem when regularizing spelling today: there are standardized spellings, such as British English or American English, that one can deploy. But in Shakespeare's period there was no such thing. In general, our solution was to use data in Early English Books Online (EEBO-TCP) to establish the most frequent spelling variant, and then use that variant as our 'standard'. However, a downside of this is that one can end up with a less than transparent regularized form, from the perspective of today's reader. An example is a clock, which was hugely more frequent than o clock in this period. We handled this issue in the dictionary through using cross-references and supplying critical information about spelling variants.

(2) Part of speech tagging and EModE. The CLAWS part-of-speech annotation system works well for present-day English (see CLAWS, n.d.), and has been adapted for Early Modern English. Rayson et al. (2007) found it to perform at 85% accuracy for Shakespearean texts. However, when a dictionary's headwords are based on a list of 'taglemmas' (i.e. lemma + POS tag), accuracy is critical. Just
one example of a problem is the word *blest*, which is variously tagged. In *which not to have been blest withall*, it is erroneously tagged as JJ (adjective). We handled this issue in two ways: we instituted a number of "fixes" to CLAWS (often simply to the lexicon), and we did a manual post-check on our core Shakespearean data.

**3) Data and genre.** Data is almost always a problem for historical corpus linguistics, because what survives is at best a reduced and patchy record of the total linguistic output of any period. Given that our project aim is to place Shakespeare in context by comparing his works with those of other writers, we needed a large quantity of comparative data. Fortunately, we have seen the advent of the transcribed 1.2 billion-word EEBO-TCP, approximately 321 million words of which span the period 1580-1640. However, this data lacks a full classification of genre. In historical work, genre is perhaps the key notion for accessing the stylistic flavour of an expression - whether it is formal, colloquial, literary, informational, and so on. Consequently, we instituted a classification scheme for the 1560-1640 period, largely based on the existing titles of works (in effect, their self-classifications).

The problems illustrated above are certainly not unique to our project. What is unique to our project is our approach to solving the over-arching problem of bringing together all the relevant information generated during these preparatory phases. Each dictionary entry is to be based on multiple pieces of information – information about spellings, part-of-speech, collocates, genre distribution, social distribution (e.g. male/female; high rank/low rank), and more – and, moreover, there are multiple information sets – Shakespeare's plays, his poetry, the Folios, the Quartos, our comparative corpus of playwrights and the EEBO-TCP. Extraction of the information is not the difficulty; we will largely be using CQPweb (Hardie, 2012) for this. The problem is more one of resources: it would take a team of researchers an inordinate amount of time, well beyond the bounds of project funding, to manually extract each piece of information and then make sense of the whole. We need a way of automatically pooling the bulk of the information, presenting it to researchers in a palatable fashion, and allowing them *in situ* to construct an interpretative summary that will constitute a dictionary entry. Our solution was to construct a database, accompanied with a user-friendly interface, for use by our team of lexicographers.

The database consists of two major parts. One part contains unchanging data, organised around taglemmas (lemma and POS tag pairing, see also above): for each taglemma identified in Shakespeare's First Folio, a series of information is automatically extracted from CQPweb and loaded into the database. This information includes overall frequency and dispersion of the taglemma in the First Folio, but also frequency and dispersion within sub-categories relevant to the First Folio, such as text genre, gender of characters, social status of characters, regularised morphological forms, and original spelling variants. Also included are overall and sub-category frequency and dispersion figures for these taglemmas in EEBO-TCP (specifically 1560-1640) and in a corpus containing plays by Shakespeare's contemporaries.
The second part of the database contains data generated by the lexicographers, including definitions, examples, cross-references and comments about their observations of the data. The user-friendly interface facilitates the generation of this content by providing access to the unchanging data stored in the database, as well as facilities for uploading manually generated data. Beyond providing access to stored data, the interface also provides other crucial functions for the project, such as helping facilitate collaboration between users (e.g. via the sharing of comments), providing version control, and helping with error- and inconsistency- checking (e.g. by providing a mechanism for selecting and updating cross-references).

To illustrate both the use of the database and interface, and the scholarly contributions of the project, we present two case studies chosen to maximize diversity:

(1) A more grammatical word: *I*. Though typically omitted from Shakespearean dictionaries, presumably on the assumption that its meaning has not changed or that it does not contribute much to understanding Shakespeare, analyses of collocates reveal that it is key in revealing character states, thoughts and feelings, as well as doing interpersonal work. It also turns out that Shakespeare had a penchant for *I*, relative to his contemporaries, at least in certain constructions, and used it to bolster particular types of characters (e.g. Desdemona in *Othello*).

(2) A more lexical word: *good*. Shakespearean dictionaries seem overwhelmed by the 2,711 instances of the word *good*, something that seems to be reflected in their widely varying accounts of the word. We will show how corpus-based analyses improve on those accounts, and actually provide support for one of the older accounts, Onions (1986/1911), in placing the usage he referred to as "conventional epithet" in pole position.

References


A longitudinal corpus-based study of *it*-extraposition constructions in Italian EFL academic writing
Erik Castello (University of Padua, Italy)

*It*-extraposition constructions have been researched extensively in various fields, including corpus linguistics (e.g. Kaltenböck 2000, 2003; Groom 2005; Thompson 2009), learner-corpus research (e.g. Römer 2009; Hasselgård 2009; Herriman & Boström Aronsson 2009), and corpus-based research aiming at providing fine-grained descriptions of lexico-grammatical patterns for English Language Teaching (e.g. Francis, Hunston & Manning 1996, 1998). These constructions are of the utmost importance to written academic English and to the field of English as a Foreign Language (EFL), as they enable writers to objectively and impersonally make statements or express opinions and ultimately develop a successful academic persona (e.g. Hewings & Hewings 2001, 2002; Larsson 2016). They can, however, pose a series of challenges to EFL learners (e.g. Hinkel 2013), which can persist even at an advanced level of linguistic proficiency, at least in the case of Italian learners (e.g. Castello 2015).

Learner-corpus research on *it*-extraposition has mainly investigated cross-sectional data, while longitudinal studies are less common. This paper attempts to fill this gap by exploring the use of *it*-extraposition constructions in academic texts written by two cohorts of Italian undergraduate language students at the University of Padua (Italy), each followed over a period of two academic years, respectively from year 1 to year 2 (138 students) and from year 2 to year 3 (73 students). This longitudinal data (189,051 words in all) is part of the Italian component of the Longitudinal Database of Learner English (LONGDALE)\(^1\). It is compared to a sub-corpus of the Louvain Corpus of Native English Essays (LOCNESS)\(^2\) (35,399 words) and to another one from the Italian component of the International Corpus of Learner English (ICLE_IT)\(^3\) (Granger, Dagneaux & Meunier 2002) (107,204 words).

The study involved the manual inspection of the concordance lines for all the instances of *it*-extraposition constructions in the corpora, which were retrieved by means of the corpus query system The Sketch Engine (Kilgarriff et al. 2014). Following Francis, Hunston & Manning (1996, 1998), the constructions were grouped according to the “pattern group” they belong to, including incorrect patterns. The following are the most recurrent patterns accompanied by examples:

1. *it* (mod) V (adv) adj: e.g. For a language learner *it is almost impossible* to completely master a second language.
2. *is* (adv) adj: e.g. I think *is absolutely possible* being able to speak a language like a native speaker.
3. *it* (mod) be n/prep: e.g. Perhaps *it would be a good thing* if people would learn it at the beginning of the process.
4. *it* (mod) be V-ed: *It can be claimed* that going abroad is a successful way to know and get in touch with foreign cultures.
5. *it* (mod) V: e.g. Moreover, *it seems that* the endangered areas are more frequently run by foreign tourism companies.

---

Subsequently, a complementary analysis was conducted into the typologies of extraposed constituents in the corpora: to-clauses, that-clauses, for/to clauses, wh-clauses, ing-clauses and erroneously *extraposed noun phrases. The following is an example of an extraposed noun phrase:

(6) *It should also be obligatory a standard set of guidelines and a formal system of accreditation.

The analysis revealed that the instances of the “it (mod) be V-ed” pattern and those of extraposed that-clauses increased from year 1 to year 3 in the LONGDALE-IT data, while those of the “it (mod) be (adv) adj” pattern and of extraposed to-clauses decreased. The paper discusses these and other related results.

The last part of the study focused specifically on the erroneous uses of it-extraposition constructions with a view to exploring the causes that bring them about. The case of the mistaken extraposition of noun phrases, in particular, was explored not only considering the role of L1 interference but also the lack of successful interplay of lexical, syntactic and discourse resources (Lozano & Mendikoetxea 2010). In order to conduct this part of the analysis, the 52 instances of extraposition of noun phrases in LONGDALE-IT and in ICLE_IT were lumped together and divided into two groups according to whether they belong to the “it (mod) be (adv) adj” pattern or to the “it (mod) be V-ed” pattern. The syntactic complexity of the sentences in which noun phrases are erroneously extraposed was measured in terms of the number of words they contain (“length of sentences”) and of the number of main and secondary clauses in them (“ranking clauses/sentence”). The complexity of the extraposed noun phrases was quantified by counting the words they consist of (“length of noun phrases”). The Mann-Whitney U test was performed to check for the statistical significance of the differences between the two groups of patterns with regard to the three variables. It turned out that the instances of the “it (mod) be V-ed” pattern with extraposed noun phrases tend to be used in longer sentences and that the differences between the values for “sentence length” in the two groups are statistically significant. These findings support the hypothesis that the combined use of it-extraposition and the impersonal passive voice - both of which are inherent features of the “it (mod) be V-ed” pattern - in long sentences is the most likely cause for the erroneous extraposition of noun phrases in the Italian data.

The paper discusses these findings as well as their implications for EFL teaching and learning.

References


1. Introduction

Research in corpus linguistics has shown that people belonging to a particular speech community have preferred ways of saying things, which are generally reflected in the use of formulaic sequences (FS), i.e. certain words (e.g. at the same time, on the other hand, for example) have “an especially strong relationship with each other in creating their meaning” (Wray 2008: 9) and play an important role in differentiating socially-situated practices (Biber et al. 2004; Hyland 2012). The predominant trend in this research area is to take what Durrant and Mathew-Aydinli (2010) call a ‘form-first’ approach (e.g. lexical bundle, n-gram), relying on the computer to identify frequent recurrent forms in a given corpus, at the expense of disregarding their structural and semantic unity (e.g. at the same, this paper we) and multifunctionality, and overlooking discontinuous units (e.g. not only….but also), to mention a few of its limitations. The present project employs a function-to-form approach in the hope of providing a pedagogically-oriented understanding of formulaic language in academic discourse. Halliday’s (2014) functional model of language, which views language as a set of systemic choices with underlying communication functions (Gledhill 2011), is useful in understanding formulaic language, which is necessary for functional language use. Drawing on Halliday’s framework of functions, this paper reports on a pilot study with the following main aims: 1) to develop an annotation scheme for a functional analysis of formulaic language in academic discourse, which can be modified and applied to other text types in future studies; 2) to shed light on possible disciplinary variation in the distribution and linguistic realisations of various functions; 3) to compare the results achieved by the traditional ‘form-first’ approach (i.e. the automatic extraction of FSs) and those by the function-first approach. The ultimate aim in this regard is to provide suggestions as to how to combine the advantages of both approaches in order to provide a full picture of formulaicity in language use.

2. Material and methods

In this pilot study, twelve texts (with a total of 19,592 words) were drawn from the British Academic Written English Corpus (BAWE), representing two broad disciplinary groupings: Art & Humanities (AH) and Physical Sciences (PS). All the texts are of the same genre (essay) and grade (D), and were written by English-speaking students in their last year of undergraduate studies or on master’s courses (years 3 and 4). FS in this study is an umbrella term covering a number of sub-categories including set phrases that are semantic opaque or grammatically irregular in varying degrees (e.g. in so far as, give way to, a great deal of), underlying frames with one or more gaps
(e.g. as $X$ as $Y$, if $X$ then $Y$), word strings that can be replaced by a single word (e.g. fail to – not, be able to – can, make a decision – decide), and formulas that are not peculiar in terms of their internal semantics/syntax, but are genre specific in the sense that they are used to realise functions in a particular type of situation (e.g. this means that, it is important to). The selected texts were manually examined to identify word sequences that satisfy at least one of the criteria mentioned above. What is more, the sequences should form a complete semantic and structural unit, whether or not the main elements are contiguous.

The identified FSs fulfil a range of functions, which fall into three broad categories within Halliday’s framework: i) ideational (or experiential) metafunction, including functions such as referring to previous research (e.g. according to), describing attributes (e.g. the length of, be responsible for), describing research procedures (e.g. tidy this up, work our way through), and manner (e.g. as a means of, by doing so, in detail, in a straightforward manner); ii) textual metafunction, including structuring signals (e.g. see for example, we can show that, as follows, in conclusion) and cohesive devices (e.g. as a result, in contrast, on the grounds that); iii) interpersonal metafunction, including evaluation (e.g. it is interesting that, play a key role in) and hedging devices (e.g. seem to, a certain degree of). In the present study, the UAM corpus tool (O’Donnell 2013) was employed for the annotation of the texts. In addition to a functional-analysis scheme, the annotation also contains information including the criterion each sequence satisfies in order to qualify as a FS and its structural make-up for the subsequent analysis.

Apart from the manual identification, the conventional form-first approach (i.e. automatic identification) was taken, using IDIOM Search (Colson 2016a), which is an online tool for the extraction of multi-word phrases, ranging from bigrams to sevengrams (see Colson 2016b for the algorithm of and improvements made by this tool in corpus-based computational phraseology). This tool is useful in capturing continuous word sequences that are not necessarily irregular or semantically and/or structurally complete, but nonetheless occur frequently enough and are often associated with a particular function in a given context.

3. Initial results

Altogether, 5480 FSs were identified in the corpus; see Table 1 for the distribution of the FSs in the two sub-corpora representing the two broad disciplinary groupings. Overall, there was an overuse of FSs in the AH sub-corpus relative to the PS sub-corpus ($p<0.0001$).

<table>
<thead>
<tr>
<th>Sub-corpus</th>
<th>No. of words</th>
<th>No. of FSs</th>
</tr>
</thead>
<tbody>
<tr>
<td>AH</td>
<td>7656</td>
<td>2288 (30%)</td>
</tr>
<tr>
<td>PS</td>
<td>11936</td>
<td>3192 (27%)</td>
</tr>
<tr>
<td>Total</td>
<td>19592</td>
<td>5480</td>
</tr>
</tbody>
</table>

Log-likelihood test: LL=16.34, $p<0.0001$
Out of the three main functional categories, the ideational metafunction was in clear dominance, accounting for 75% of all FSs, whereas textual functions made up 14% and interpersonal 11%. Within the ideational category, the function of manner stood out (11%), followed by the functions of describing attributes (9%) and argumentation (8%). The textual category was dominated by structuring signals (location, sequencing, introductory remarks, presenting results), which accounted for 41% of all textual functions, followed by cohesive devices (23%). Out of the interpersonal category, 33% were of the evaluation function, and the other two functions that may be interesting to look further into were obligation (12%, e.g. *it is necessary to, ought to, have to*) and personal opinion (7%, e.g. *in my view, I think*).

Table 2 presents the distribution of the three categories in the two sub-corpora. Again, significant difference was found between the two broad disciplinary groupings ($p<0.001$). Most noticeably, the text-oriented functions seemed to be employed to a particularly large extent by the students of PS. A closer look at the distribution of various functions of this category revealed that the main difference between the two sub-corpora rests with the functions related to cohesive devices.

<table>
<thead>
<tr>
<th>Sub-corporus</th>
<th>Ideational</th>
<th>Textual</th>
<th>Interpersonal</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>AH</td>
<td>1807 (79%)</td>
<td>208 (9%)</td>
<td>273 (12%)</td>
<td>2288</td>
</tr>
<tr>
<td>PS</td>
<td>2313 (72%)</td>
<td>568 (18%)</td>
<td>311 (10%)</td>
<td>3192</td>
</tr>
<tr>
<td>Total</td>
<td>4120</td>
<td>776</td>
<td>584</td>
<td>5480</td>
</tr>
</tbody>
</table>

$df=2, \chi^2=84.808, p<0.001$

Comparing the figure resulted from manual identification with that of IDIOM Search, the latter retrieved fewer FSs (3888 units). Among those manually identified FSs, 30% are identical with those automatically retrieved, 29% are partly compatible and 41% failed to be captured by IDIOM Search, indicating a need to combine the two approaches in the study of formulaicity.

### 4. Concluding remarks

In conclusion, the initial results are significant enough to warrant further investigation. Among others, a qualitative analysis will be carried out to look at the few functions that stood out in the overall comparison above and to find out the particular discipline(s) where they tended to be employed. The results will be compared with those of previous studies (e.g. Hyland 2008) that employ the ‘form-first’ approach. Given the small size of the corpus, individual differences will also be brought up in the discussion. This pilot study will be expanded later on by including more texts from BAWE to provide more illuminating insights into disciplinary variation. In addition, I will include a selection of published research articles as well as essays written by L2 students, who are of the same academic level as those L1 students in the pilot study. Through a comparative analysis of student essays and published research articles, the study will shed light on the degree of formulaicity in
producing academic discourse across disciplines, and bring out similarities and/or differences between novice and expert writing, with pedagogical implications for the training of novice writers in scientific fields.

References


“Japanese English”: A Descriptive Grammar of Educated Written English in Japan
Kolawole Olagboyega (Yamanashi Gakuin University, Japan)

Arguments rejecting the pedagogic notion of “Standard English” or “correctness”, and suggesting that all forms of English are equal, have resulted in the proliferation of terms such as “Indian English”, “Singapore English”, “Filipino English”, “Nigerian English” etc., which are claimed to be on precisely the same equal footing with “American English”, “British English”, “Australian English” (Coleman 1987:13; Kachru 1986a, 2005; Kortmann, et. al., 2012).


Kachru (2005), identifies six aspects of Japanese English from the “historical”, “functional”, “formal”, “attitudinal”, “pragmatic”, and “acquisitional” perspectives. However, the most comprehensive study on the subject of “Japanese English” was conducted by James Stanlaw (2004). In his seminal study done from ‘an anthropological linguistic perspective’, Stanlaw describes “Japanese English” as ‘a created-in-Japan variety for use by Japanese in Japan regardless of how they may appear to native English speakers’. In other words, the Japanese do not aspire to approximate the native norm. He claims that in the larger context of ‘world Englishes’, “Japanese English” is so entrenched that English has become ‘…a Japanese language’. This is an extreme position which few scholars share with Stanlaw.

Meanwhile, the tendency among researchers on the subject of English in Japan has largely been to provide glossaries of coinages and other lexical modifications, and the listing of isolated examples of divergence, and present them as “the features” of “Japanese English” (cf. Stanlaw 2004). Caught helplessly in this controversy especially in a country such as Japan, where English language is chiefly acquired through formal education, is the classroom teacher, who needs to know what form of English to teach, and which reference books to use.

This study does not select isolated examples of forms to corroborate or falsify any theoretical position or construct, which has been the general trend of research in the field. Instead, it seeks to provide a descriptive grammar of aspects of educated written English in Japan, on which those concerned with teaching English in Japan particularly at junior high, high school and university can draw; it seeks to demonstrate that across the range of forms which are regularly identified as “errors” in the English written by educated Japanese, there are some environments which regularly reflect “Standard English practice” and others where “divergent forms” are manifested with some degrees of frequency. The discrimination between the different types of environment gives some idea of the possible reasons for this variation and how to set about correcting it in the classroom.

As an attempt to conflate the existing pedagogical concept of "Standard English" and the emerging theoretical notion of "standard non-native varieties of English" (Milroy and Milroy 1987; Quirk 1989), this study looks at the stability of the
claimed "characteristic" forms of "Japanese English" and shows the statistical likelihood of their occurrence in particular syntactic and semantic environments. This particular study focuses on the realization of the grammatical categories that are typically associated with the constituents of the noun phrase (NP).

This approach is both pedagogically and theoretically interesting inasmuch as it identifies the divergent forms. The classroom teacher, for example, may know what to "correct" and the textbook writer what to highlight. The theoretical linguist who argues for the existence of non-native standard varieties of English (cf. Todd and Hancock 1986; Williams 1987) has also got ready evidence on which to draw; evidence that can also validate the concept of "fossilization" (cf. Selinker 1972), which seeks to account for the adult non-native speaker’s grammatical variability.

The corpus consists of material that appears in the four Japanese national English-language newspapers, Asahi Evening News, Japan Times, Mainichi Daily News, and The Daily Yomiuri which comprises the editorials, articles, advertisements, letters to the editor, etc.; government publications, such as those of the Japanese Ministry of Education (MEXT). The data used covered a period of 15 years (2000 to 2015). The corpus also includes articles published in English by Japanese University professors. Statistical information is given in the text itself. Because we are interested in the language produced by a people or group of speakers rather than the individual variability within the group, the database is cross-sectional rather than longitudinal.

The newspapers have been selected for this study for several reasons. Published daily (Monday to Sunday), they are the most widely circulated national English-language newspapers in Japan read by the whole public, whose proficiency in English ranges from the lowest to the highest (native-like). The newspapers constitute what might be called the Japanese quality press.

Since the corpus also includes data collected from the articles published in English by the Japanese university professors, and data collected from MEXT publications as well as the English of Japanese newspapers, we are associating these with "educated English" (call it the "acrolectal") in this study. These are the highest levels at which we are readily able to find a corpus from the daily communicative experience of the people, large enough to be well representative of the major forms, and quite convenient for detailed examination. Other bases, such as students' writing collected at various levels, will show tendencies that are generally associated with early and middle learners (the "basilect" and the "mesolect" speakers). It is necessary to emphasize tendencies in relation to a database because there seems to be no objective way of dividing the cline of bilingualism. In sociolinguistic terms (Kirkpatrick 2005; Bolinger and Sears 1981; Magura 1985), the levels of proficiency are group into those broad stages of the acquisition process, each of which is associated with a variety of the language.

The data for this study was collected manually, and was therefore very laborious. Each detected divergent form is then manually fed into the Word document, which serves as the computerized "tool" and corpus for the study.

As we are interested in the description of data rather than the explanation of a theory or process, the approach is more inductive than deductive. The realization of the grammatical categories that are typically associated with the constituents of the noun phrase (NP), are examined in the English of the newspapers, in the writings of university professors, the government publications and the writings of
university students. The environments where persistent patterns and tendencies emerge are described and tabulated with a view to determining the extent to which the patterns may be said to represent stabilized usage based on a specifically Japanese syntax and semantics as opposed to (American) Standard English practice.

For example, our corpus suggests that Japanese writers of English tend to omit the definite article in an NP where its presence is described as being customary - i.e., the head of such an NP is said to be customarily preceded by the definite article. Such nouns include names of: local, national and international bodies, public facilities - hotels, hospitals, restaurants, cinemas, etc.; certain countries, universities and schools, newspapers, etc. (cf. Quirk et al 1985:289, 296). Out of 1080 such NPs counted in our corpus, the article was omitted in 650 (60% of) cases:

1. Mr. X ... has flown to United States to undertake a management development course ....
2. The national baseball competition ... enters its third day today with matches at University of Tokyo ... .
3. Also on the plane is Mr. X, senior staff writer, who will cover the match for Daily Yomiuri.
4. Daily Yomiuri has not reported the news.
5. He described the conditions at Akita University Teaching Hospital as appalling.

In the discrimination between different types of syntactic environment, our pedagogical aims will take precedence, and we shall be suggesting one grammatical approach or another. The main purpose, however, is not to prescribe any particular approach but to furnish the teacher-trainer, textbook writer and curriculum designer with an eclectic mixture of methodological frameworks which will be useful in approaching a particular problem.

The relations between the standard forms and the divergent forms, and their percentages of co-occurrence will provide helpful insights into various theoretical issues. For example, the corpus shows no grammatical categories that regularly occur divergently only and never standardly. If we accept the general view that there is a distinctive Japanese English usage that can be clearly distinguished from standard practice in terms of such tendencies as "omission of articles, pluralization of non-count nouns, etc.", then we must allow for a great deal of overlap between "Japanese English usage" and standard practice in the language produced by educated Japanese. It will be demonstrated that educated Japanese, for instance, do not consciously omit articles in every context where standard practice would require them, nor are non-count nouns consistently made to take the regular plural morph whenever they are expected to have semantically plural interpretations.

References


Corpus Pragmatics and corpus-based discourse studies are becoming increasingly important sub-disciplines of Corpus Linguistics, as e.g. evidenced through publications like Aijmer & Rühlemann (2014a), the Yearbook of Corpus Linguistics and Pragmatics series edited by Romero-Trillo, or Baker & McEnery (2015). Yet, in order to investigate meaning in context, most of the research described in these publications still relies on more traditional and basic corpus linguistics methodology – i.e. what Aijmer & Rühlemann (2014b: 8) refer to as the “vertical reading” of concordance lines – which is only suited for very limited, small-scale pragmatic analysis, or on largely unsuitable or imprecise techniques, such as simple keyword or frequency analyses. At least part of the reason for this shortcoming probably lies in what Searle (1963: 136ff.) refers to as “[t]he speech act fallacy”, i.e. the mistaken belief that single words may allow us to characterise and/or identify meaning adequately. However, prior attempts to resolve this issue in order to be able to identify contextual meaning, such as basic collocation analysis or even the identification of “functional profiles” (Adolphs 2008: 10) still have not advanced the field of corpus pragmatics enough to make the large-scale analysis of pragmatic meaning possible, not only because they are too limited in scope, but also because they do not leave a clear record of the facts since they do not – as yet – “[...] make explicit the relationship between individual speech act expressions and their distribution across different con-texts”, as Adolphs (ibid.) stipulates needs to be done. The only way in which such an endeavour can be realised is to drive forward the creation of pragmatically annotated corpora, and with it, the methodology required for achieving this.

As illustrated in Weisser 2014 & 2017, the first version of the Dialogue Annotation and Research Tool (DART) already presented a major novel way of enriching dialogue data largely automatically with pragmatically-relevant annotations on a number of different levels, thereby taking the potential for genuine corpus-based approaches to the field of pragmatics one step further. The distinct levels covered there comprise syntax (both traditional and extended ‘sentence’ types), semantics (‘topics’), semantico-pragmatics (‘IFIDs’; Searle 1969: 16), surface polarity, and pragmatics (in the form of speech acts). The number of potential individual speech acts the first version was able to recognise with a high degree of precision (cf. Weisser 2016a) was 57, some of which could occur in combination. This number already exceeded that of the speech acts employed in most traditional taxonomies, such as those established by Austin (5), Searle (5), as well as those derived from the latter for the annotation of the SPICE Ireland (9; Kallen & Kirk 2012), by far. Even in comparison to the more practice-oriented taxonomies employed in recent NLP-oriented projects, such as the Maptask Corpus (12; Kowtko et al. 1993), DAMSL (31; Allen & Core 1997), or Switchboard DAMSL’s “approximately 60 basic tags” (Jurafsky et al. 1997: 1), DART 1.0 already performed rather well, too, especially as the taxonomies implemented there mostly still needed
to be applied manually before allowing computational linguists to devise more or less successful algorithms based on machine learning techniques.

Version 2.0 of DART now supports an even more fine-grained basic taxonomy of more than 120 basic categories and their potential combinations, distinguishing between different types of speech acts as realised through and in different c-unit types, the sequencing of units in dialogue, the influence of modality, polarity, etc. In comparison to the first version, it also features a more robust grammar for recognising different syntactic types, a larger inventory of IFIDs, and an improved inferencing mechanism for deducing speech acts, all based on symbolic, rather than probabilistic identification strategies. The annotations produced in DART thus not only make it possible to achieve the aims pointed out by Adolphs, but also make it possible to carry out further investigations into the form–function relationship embodied in, and expressed through, the different levels, potentially leading to far deeper in-sights into the mechanisms that underlie different communicative strategies, as already illustrated to some extent in Weisser (2016b), where the interactional behaviour of one British and one American call-centre agent was profiled one against the other, as well as against that of their respective callers.

In this talk, I first want to present the design of the new version of DART in terms of the enhancements in its interface and corpus handling features compared to the earlier version. This will then be followed by a brief illustration of the annotation process and analysis options, finally pointing forward to how these features can be exploited for various purposes in research into Corpus Pragmatics.

References


A linguistic typology of American television programs

Tony Berber Sardinha (São Paulo Catholic University, Brazil) and
Marcia Veirano Pinto (São Paulo Federal University, Brazil)

Over the years, television programs have been classified in a large number of different ways, ranging from broad categories reflecting the general topic of the program (e.g. shows about cooking, politics, cars, etc.), the target audience (e.g. children’s/teenagers’/women’s shows, etc.), the time of showing (morning/late night/Sunday morning shows, etc.), to more specific taxonomies based on the perceived genre of the show (soap opera, news, talk show, etc.) (Creeber, 2008; Frank, Becknell, & Clokey, 1971; Mittell, 2004; Rose, 1985; Wasko, 2010). Both general and specific taxonomies are in use today, in the television industry and in academia, to refer to groups of shows that share common features. However, to date no classification scheme has been developed that relied primarily on the linguistic features of the shows as a basis for the taxonomy. The goal of the current study is exactly to develop such a linguistic taxonomy of the verbal language of television programs shown in the United States. Our analysis is restricted to the spoken component of the television programs. An analysis of the visual and sound components would require a different method and would probably yield different results. Existing research on the verbal language of television from a corpus perspective has focused on comparing selected television registers among themselves or contrasting particular television programs to naturally occurring conversation (Al-Surmi, 2012; Bednarek, 2010, 2011, 2012; Csomay & Petrovic, 2012; Quaglio, 2009). No previous research has proposed a corpus-based linguistic taxonomy of the spoken language of television programs.

Our taxonomy has been developed from a multi-dimensional (MD) corpus-based perspective, using the dimensions of variation across American television registers uncovered by Berber Sardinha & Veirano Pinto (2014a; forthcoming). The MD framework is a corpus-based method introduced by Biber (Biber, 1988 et seq.; Berber Sardinha & Veirano Pinto, 2014b), whose goal is to identify the underlying parameters of variation among texts (the ‘dimensions’). The dimensions are based on groupings of correlated linguistic characteristics. These groupings in turn are identified through a series of factor analyses of the normed counts of hundreds of linguistic features found across the texts (cf. Friginal & Hardy, 2014). The corpus employed for this analysis was the USTV corpus, consisting of 31 registers (programs), totaling 5.3 million words. The corpus was carefully designed so as to represent the multitude of programs presented on contemporary American television (terrestrial and cable). In addition, the size of each corpus section was calibrated so as to reflect the inherent linguistic variation among the texts, following Biber’s (1993) proposal for corpus representativeness (cf. Berber Sardinha, 2014). As such, a pilot version of the corpus was collected, cleaned up, hand-checked and tagged for part of speech using the Biber tagger. The variation across the texts in each register was then assessed through a preliminary MD analysis, and extra texts were allocated to the registers that exhibited more variation. The final version of the corpus was then tagged with the Biber Tagger, and the counts of nearly 200 characteristics were taken with the Biber Tag Count program. The normed counts were analyzed factorially, thereby identifying four factors, which were interpreted as dimensions of variation, namely: 1. Exposition and discussion vs. Simplified interaction; 2. Simulated involvement; 3. Recount; 4. Engaging presentation. The dimensions captured the majority of variation among the registers, namely 80.3% (dim. 1), 63.6% (dim. 2), 72.3% (dim. 3), and 55.4% (dim. 4). Each text in the corpus was scored on each of the four dimensions. The scores were obtained by adding the standardized frequencies of the features that loaded on the
positive pole of each factor and by subtracting the features that loaded on the negative pole from the previous sum.

The linguistic typology was based on a cluster analysis of the dimension scores of each text, following Biber’s (1989) proposal for text type identification (see also Berber Sardinha, forthcoming). In an MD text typology, texts types are ‘[g]roupings of text that are similar in their linguistic form’ (Biber, 1989: 13). Text types are determined through cluster analysis, which:

groups texts such that the texts within each cluster are maximally similar to each other in their exploitation of the textual dimensions, while each cluster is maximally distinct from the others. That is, those texts with the most similar dimension scores are grouped in each cluster. (Biber, 1989: 13)

A cluster analysis was performed on the dimension scores in SAS University Edition using the FASTCLUS procedure, which yielded disjoint clusters. Disjoint clusters were preferred as ‘there was no theoretical reason to expect a hierarchical structure’ in the text typology (Biber, 1989: 42). A challenge in cluster analysis is the determination of the optimal number of clusters in the data. In previous research of this kind, the Cubic Clustering Criterion statistic provided by the FASTCLUS procedure was used to ‘provide a measure of the similarities among texts within each cluster in relation to the differences between the cluster’ (Biber, 1989: 42). These heuristic devices ‘reflect goodness-of-fit: the extent to which the texts within a cluster are similar, while the clusters are maximally distinguished.’ (Biber & Kurjian, 2007: 120). An examination of the values of the CCC statistic seemed to indicate the presence of nine clusters in the data. A provisional extraction of six clusters was then conducted. The texts in each cluster were distinguished with respect to the distance from the cluster centroid (Biber, 1989: 42). Core texts include more of the salient features on the cluster, whereas peripheral texts display fewer of the major characteristics of the cluster, which makes them ‘relatively dissimilar to the central cluster characterization, but even more dissimilar to other clusters.’ (Biber, 1989: 16). The clusters were interpreted qualitatively by considering how the major linguistic features of the different dimensions were used in the texts, in addition to the mean scores of the cluster on each dimension as well as the major registers included in the cluster. This linguistic profile was used to characterize the individual clusters as linguistic text types.

As mentioned, nine provisional clusters have been identified, which have the following major characteristics. Cluster 1 includes texts that are extremely expository and highly non-interactive. It is comprised mostly of news debate programs, live politics broadcasts and newscasts. Cluster 2 is in some ways similar to cluster 1, in that it comprises texts that are very highly expository, and moderately non-interactive, but retrospective. The major registers in the cluster are investigative news programs and talk shows. Cluster 3 is very highly expository and moderately retrospective. The texts are predominantly live broadcasts from the US Congress and news debate shows. Cluster 4 is very highly non-interactive, non-retrospective and factual; the most typical registers are sports broadcasts and commercials. Cluster 5 is simplified, interactive and moderately retrospective. The most frequent registers are reality shows and legal programs. Cluster 6 is simplified, highly non-interactive, very highly non-retrospective and highly engaging. The most typical registers are culinary and lifestyle shows. Cluster 7 is highly simplified, moderately non-interactive, non-retrospective and highly fact-based. Its most common registers are children’s and teenagers’ series. Cluster 8 is very highly simplified, and highly interactive, and the most typical registers are preschoolers’ shows, and children’s animations. Finally, cluster 9 is very highly simplified, very highly interactive, highly
retrospective, and moderately engaging. The most typical registers are soap operas. As mentioned, this is analysis is not final, and the text type labels are provisional; we intend to look at ways of improving it before a final taxonomy is reached. Overall, this working typology of American television programs differs from previous taxonomies with respect to both the number of types identified and the categories determined. Furthermore, to the best of our knowledge, this is the first linguistic typology of television programs, and one of the few MD taxonomies of texts of any kind developed so far. In the paper presentation, examples of each cluster will be provided, in addition to interpretive labels of the clusters and a detailed discussion of the results.

References


Frigival, E., & Hardy, J. A. (2014). Conducting Multi-Dimensional analysis using SPSS. In T. Berber Sardinha & M. Veirano Pinto (Eds.), *Multi-Dimensional Analysis, 25 Years*
This paper explores the potential of key word analysis to shed light on lexical cohesion in literary texts, showing how key words can support the study of cohesive networks. Through the analysis of H. P. Lovecraft’s *At the Mountains of the Madness*, it explores how a corpus linguistic approach can further our understanding of lexical cohesion as an incremental textual feature that contributes to literary meaning.

Key word analysis has been extensively used for the study of literary texts (for example, Stubbs 2001, Stubbs 2005, Toolan 2009, Culpeper 2009, Walker 2010, Mahlberg & McIntyre 2011). However, an aspect of key words that does not seem to have been explored yet is their role in building cohesion: key words, as a form of repetition, contribute cohesion to a text. Moreover, considering key words as items that occur repeatedly as the text goes on also emphasises the directionality of the repetition and the fact that meaning is incremental, reflecting the progression of the text. In contrast, in corpus linguistics the emphasis is on the amount of data and meaning is discussed from a cumulative, rather than directional, point of view. With this paper we aim therefore to explore the potential of key words to act as nodes of cohesive networks that develop incrementally throughout a text. As Thornbury (2010: 279) explains, “while a list of keywords is not in itself a semantic network, it provides the raw data out of which such a network can be constructed”.

In order to test this potential, we carry out a study of cohesive networks in *At the Mountains of the Madness*. First, we use a log-likelihood test ($p \leq 0.0000001$) to generate a list of key words comparing the text (41,431 words) with a reference corpus of coeval American fiction (over 2M words). We then select key words to examine in detail as potential nodes of cohesive networks. Two interrelated criteria are used for the selection of the key words: (i) frequency – the selected key words must occur frequently throughout the text, and (ii) thematic relevance – the selected key words must relate to the themes of the short novel. Frequency is important because a word that is repeated frequently across a text contributes importantly to the cohesion of that text. Thematic relevance is important because themes and “aboutness” are textual features that develop incrementally throughout the text. A word that is both thematically relevant and frequent, as a content key word can be, is therefore a possible candidate to represent a node of a cohesive network. For the purpose of exemplifying the method of cohesive networks identification, we focus on two key words that meet both the aforementioned criteria, *mountains* (48 occurrences) and *peaks* (29 occurrences), using traditional literary criticism to inform our understanding of what is thematically relevant in the text.

With the help of the concordancer, we analyse the key words’ patterns and their local textual functions (Mahlberg 2005, 2006). We identify two textual behaviours that are shared by both key words: a semantic preference of “Height”, as seen in great, highest, steep, titan, elevation, gigantic, greatest, loftiest, mighty, etc., and a negative semantic prosody referred to as “Mysteriousness”, that characterises *mountains* and *peaks* as a frightfully mysterious place, as seen in the co-occurrence of words such as madness, horrible, evil, nameless, terrible, unknown, brooding, cryptic, forbidden, frightful, horror, looming, mysterious, mystery, etc. These patterns are built incrementally with every occurrence of the key words: when *mountains* and *peaks* appear, they co-occur with the words that form the shared semantic preference/prosody. Each occurrence of these key words thus connect together the instantiations of the semantic preference/prosody, creating a net of connections that builds incrementally and spans across the whole text. These connections are conceptualised as a
cohesive network based on the repetition of the key words and the semantic preference/prosody. Figure 1 below shows an example of the network. The figure shows 12 random sentences in which *peaks* and *mountains* occur, sorted in the order they appear in the text. The nodes of the network (the key words) are underlined, while the words belonging to the shared semantic preference and prosody are in bold. Every successive instantiation of the semantic preference/prosody builds on the previous ones, through the link that the repetition of the key words establishes.

(1) The last lap of the voyage was vivid and fancy-stirring, **great barren peaks** of *mystery looming* up constantly against the west as the low northern sun of noon or the still lower horizon-grazing southern sun of midnight poured its hazy reddish rays over the white snow [...]. (2) [...] since an aeroplane survey of the nearly exposed rock surfaces showed an entire absence of those Archaean and primordial strata for which he was looking, and which formed so great a part of the **colossal peaks** that **loomed** up at a tantalising distance from the camp. (3) [...] but at this time and place, with those **dark, unknown mountain peaks soaring stupendously** ahead, that anomalous elder-world discovery in our minds, and the pall of probable disaster enveloping the greater part of our expedition, we all seemed to find in it a taint of latent malignity and infinitely evil portent. (4) The **unknown mountains** ahead rose **dizzily** up like a fearsome rampart of giants [...]. (5) One had to be careful of one's imagination in the lee of those **overshadowing mountains of madness**. (6) In spite of all the prevailing horrors we were left with enough sheer scientific zeal and adventurousness to wonder about the **unknown realm beyond those mysterious mountains**. (7) As we drew near the **forbidding peaks**, **dark and sinister** above the line of crevasse-riven snow and interstitial glaciers, we noticed more and more the curiously regular formations clinging to the slopes; [...] (8) The touch of **evil mystery** in these barrier mountains, and in the beckoning sea of opalescent sky glimpsed betwixt their summits, was a highly subtle and attenuated matter not to be explained in literal words. (9) [...] a **frightful** line of **peaks** had shot suddenly up amidst the most appalling din and chaos—and earth had received her **loftiest** and most **terrible mountains**. (10) This vast nighted gulf had undoubtedly been torn by the great river which flowed down from the **nameless and horrible westward mountains**, [...]. (11) Perhaps we were mad—for have I not said those **horrible peaks were mountains of madness**? (12) For this far violet line could be nothing else than the **terrible mountains** of the **forbidden land**—**highest** of earth's **peaks** and focus of earth's evil; [...].

**Figure 1.** Example of cohesive network

In Sentence (1), *peaks* occurs with **great, barren, mystery, and looming**. When the reader finds *peaks* again in the text in Sentence (2), a connection is established with the previous occurrence. In this case, *peaks* occurs with **colossal and loomed**, which relate to **great and looming** respectively. In Sentence (4), *mountains* recalls the previous uses of *peaks*; as both terms refer to the same fictional place. Here, the adjective unknown links back to the earlier occurrence of the same word (in Sentence (3)) or to the use of related items (*mystery, in Sentence (1), for example). These connections are established every time *peaks* and *mountains* occur, and every occurrence reinforces the link in an incremental fashion. In Sentence (9), towards the end of the short story, *peaks* occurs with **frightful**, while *mountains*
occurs with loftiest and terrible. Frightful and terrible refer back to all of the occurrences of the “Mysteriousness” semantic prosody, whereas loftiest builds on the repetition of the “Height” semantic preference.

We argue that this network, in addition to creating lexical cohesion in the text, participates in the process of building up the fictional world. The representation of mountains and peaks is in fact cohesive and consistent throughout the text, and this contributes to define the spatial locales of At the Mountains of Madness. As recognised by many literary critics (e.g. Ralickas 2007, Kneale 2006), spaces and places in this short novel are a fundamental aspect to establish the atmosphere of malevolence and awe that characterises the text.

Overall, building on corpus linguistic models, cohesion can be conceptualised as the sum of relationships between lexical items rather than between individual words. The approach we present in this paper shows how it is possible to study cohesion as a feature that spans across the whole of a text. We argue that key words can function as nodes of cohesive networks, and that their repetition contributes incrementally to the construction of literary meanings.

References


Do You Also See what I See: Russian-Ukrainian Conflict in European, Ukrainian and Russian Media
Anastasiia Kryzhanivska (Bowling Green State University, OH, USA)

In war or conflict discourse the power of media is crucial because it affects the image of the country on the international stage as well as public mood and atmosphere in the country. The Russian-Ukrainian conflict has been a matter of debate between many countries including but not limited to Russia and Ukraine. This research may offer some answers for the international community regarding the tension between Russia and Ukraine.

By focusing on the Russian-Ukrainian conflict, the key actors of the conflict and attitude to them is presented by looking at the most frequent words in op-eds of different media sources (European – The Guardian, Ukrainian – TSN, and Russian – RT). The research was conducted as an attempt to bridge the gap between conflict media representation theory and the actual linguistic impact of news articles on the reader by answering the following questions:

- What are the most frequent words in op-eds in relationship to Russian-Ukrainian conflict in European, Ukrainian and Russian media?
- Who are the key actors of Ukrainian-Russian conflict from the point of view of different media sources?

Methodology

The current study is based on the methodology of comparative analysis of Kutter and Kantner (2012) who looked at the collocates of the most frequent lexical items in news war reporting, and Jorge (2014) who investigated the image of Ukrainians in media. For the study three mini-corpora were created consisting of op-ed articles published in August 2014 – March 2015 in three news sources: European – The Guardian, Ukrainian – TSN, and Russian – RT. In total, the corpus consists of 68 articles with 62,319 word tokens (The Guardian – 24 articles, 20,929 words; TSN – 23 articles, 20,647 words; RT – 21 article, 20,743 words). The search function and key words were used to select the articles.

The articles were gathered from three websites: European – The Guardian (http://www.theguardian.com/uk), Ukrainian - TSN (http://tsn.ua/) and Russian – RT (http://russian.rt.com/). A concordancer (AntConc) was used to organize lexical items and their co-text into concordance lines. Excel spreadsheet was used record the results of articles analysis.

Results and Discussion

After compiling the corpus of 68 op-ed articles in total, the most frequent words were found using the Word List menu option in AntConc. Lexical items with the highest frequency of occurrences in The Guardian, TSN and RT can be seen in the graph below:
The above graph shows that the most frequent lexical items in *The Guardian* and in *TSN* coincide: *Ukraine*, *Russia*, *Putin*, *Russian*, *war*. However, *RT* suggests a slightly different sequence: *Ukraine*, *Russia*, *country*, *year/years*, *people*. Obviously, *Ukraine* and *Russia* are considered to be key actors in all three sources, even though the number of occurrences of these words is not the same. *Russia* is the most frequently used lexical item in Ukrainian source *TSN*; adjective *Russian* takes the second place. In *RT*, however, *Ukraine* is the most frequently mentioned word. For *The Guardian*, *Ukraine*, *Russia* and *Putin* have approximately the same frequency. This analysis reveals that from the Ukrainian media perspective, *Russia*’s actions deserve the most attention. From the Russian perspective, on the contrary, *Ukraine* is the one that has to be mentioned the most. And *The Guardian*, as the third party, equally describes *Ukraine*, *Russia* and Russian representative *Putin*.

Moreover, this comparison suggests that *RT* does not use the word *war* very often, unlike *The Guardian* and *TSN* where it is used at least three times more often. The reason for this big difference in frequency of occurrences might be the fact that Russian-Ukrainian conflict is not considered as war in Russia. It is usually viewed as misunderstanding between the *countries* and *people*. This point of view is also reflected in the frequency of use of these two lexical items. Similarly to Chen (2013), it was assumed that rare mentioning of *government* tells about the government’s ineffective role in the conflict. Instead, the articles use personal names (places, people, etc.) to particularize the actions. In addition, instead of covering only the current state of affairs in the country, the Russian media tends to recollect events happened in the past *year/years*. Assumptions about future events can be made based on the previous experiences and conflicts between Russia and other countries.

Three mini-corpora were also analyzed in terms of the proper names (names, places, countries, capitals) that are expected to appear in op-eds due to their immediate relevance to the Russian-Ukrainian conflict. The following graph compares the use of personal names across three mini-corpora:
The above graph shows that each media source represents the Russian-Ukrainian conflict in relation to respective country. So that, European *The Guardian* sees *Europe* among active participants, Ukrainian *TSN* considers *Crimea* to be crucial in the conflict, and Russian RT has *Moscow* in top three actors. The more careful investigation of data, however, reveals certain difference between how the personal names are used in different contexts.

The most frequent personal name in *The Guardian* and *TSN* is *Putin*. The prevalent number of its occurrences tells that he is one of the key actors in Russian-Ukrainian conflict. *RT*’s most frequent personal name is *USA*. The media represents the US as Ukrainian ally who is going to provide it with weapons and send humanitarian support. Even though the American government declined Ukrainian appeal for help, the mere intention to do that in the past brought the United States into the foreground of *RT*’s discussion.

*Kiev* and *Moscow* are both very frequent personal names in all three corpora. Interestingly, *Kremlin* is sometimes used as a substitute of *Moscow* and Russian government in general in *TSN*. The occurrence of this lexical item is twice more frequent in Ukrainian source than in *The Guardian*, and at least three times more frequent than in *RT*. Indeed, referring to either Russian president *Putin* or Russian government in *Kremlin* is not very common in *RT*. It might be explained by the overall shifted focus of Russian-Ukrainian conflict in Russian media so that the attention is paid to *Europe*, *USA* and *Kiev* instead of to the actions of the Russian government or *Moscow*.

*Poroshenko* and *Obama* are almost equally mentioned across all three corpora even though only one of them is directly connected to the ongoing conflict. It could be attributed to the fact that *Poroshenko*’s presidency term has recently started at that time, which does not make him responsible for the decisions made by previous government. *Kiev* instead is a more general term to be used referring to the Ukrainian government.
An interesting correlation between Donetsk, Luhansk, and Donbas can be seen across three corpora. Only Ukrainian TSN acknowledges that the ongoing conflict concerns the whole Donbas region, and not only particular parts of it, such as Donetsk and Luhansk. It tells about the general awareness about the real situation in Eastern Ukraine. European and Russian media seems to separate Donetsk and Luhansk and single them out from other parts of Donbas region, and hence – limiting the conflict to only those two cities. In reality, every village, town and city in close distance to Donetsk and Luhansk is involved.

Germany in The Guardian and RT is mentioned almost equally. The co-text of this lexical item is very different, though. The Guardian nominates Germany as the most influential country representing the European Union. But RT describes it as Ukrainian ally along with the USA. Even though German government expressed its sincere concern about the situation in Ukraine, no significant action has been undertaken. The Guardian compared Putin to a historical figure of Hitler, saying that their methods of taking over the territory are very similar. However common this comparison in Ukraine is, only one occurrence of Hitler was found in TSN. At the same time, RT mentions Hitler as a response to the comparison of Putin to Hitler, usually impersonalizing and passivizing the accusation (it is said that…; Putin is compared to...).

Conclusions

Based on the above results, it appears that the most frequent lexical items with regards to the Russian-Ukrainian conflict in op-eds in The Guardian and TSN are: Ukraine, Russia, Putin, Russian, and war. The most frequent words in RT are: Ukraine, Russia, people, country, and year/years.

Even though three mini-corpora included only 62,319 word tokens in 68 articles from three media sources, the results show the difference of the conflict media coverage in different countries. The Guardian blames Putin, the Russian paper blames the USA, and TSN emphasizes the hardships of the current situation in Ukraine and highlights the war conflict.

References


Understanding street harassment of children: identifying recurrent behaviours in a corpus of young people’s accounts of harassment

David Wright, Lucy Betts, Rachel Harding, Catarina Sjolin Knight, Sheine Peart and Kendall Newbold (Nottingham Trent University, UK)

Background

This paper reports findings of a multidisciplinary pilot study investigating children’s experiences of street harassment in England. Combining expertise in linguistics, psychology, social work, law and education, the project explores the nature, scale, and impact of children’s experiences of harassment as self-reported via a bespoke web-based app. The term ‘harassment’ is used by the Crown Prosecution Service in England and Wales to refer to offences which ‘cause alarm or distress’ or ‘put people in fear of violence’, and although reported anecdotally, street harassment is not officially recorded in crime statistics for children (ONS, 2015). Similarly, while there has been research attention paid to ‘stranger harassment’ and sexual harassment of adults (e.g. Magley, 2002; Fairchild and Rudman, 2008; Hlavka, 2014), there is very little existing research which focuses on children as the targets of harassment. This paper, therefore, uses a corpus linguistic method to analyse the ways in which children describe the harassment they have experienced, identifying trends and patterns of incidents, and shedding light on how children perceive street harassment and the resources they have to cope with it.

Data collection and the corpus

A bespoke web-based app was designed to collect children’s reports of street harassment incidents, which they could access from their computers, smart phones and tablets. Children were given a unique reference number to preserve anonymity but to allow for the follow-up of serious causes for concern. The app required respondents to select from a number of options for ‘what happened’, for example ‘I was stared at’, ‘I was pushed/hit’, ‘I was beeped at with a car horn’, and choose from a number of emotions to express how the event made them feel, such as ‘nervous’, ‘afraid’, ‘proud’ (adapted from Ebesutani et al., 2012). In addition, children were asked to select whether they were alone, in pairs or in a group and where they were when the incident(s) took place, as well as provide details of their age, gender and ethnicity. Following these closed-choice questions, children were invited to provide their own narrative account of what happened, as the app asked: ‘If you would like to tell us more about what happened and how you felt, please use your own words to describe the incident’. It is these free-text comments that comprise the corpus data for this paper.

Data was collected from pupils from two schools in the East Midlands and one in London, following appropriate ethical approval. The children in this work are aged between 11 and 16 years of age, which is younger than previous research in harassment of young people (e.g. Swim et al., 2003). A total of 115 reports across the three schools were submitted. Of these, 61 included free text comments – 21
from boys, 38 from girls, and 2 from respondents who did not disclose their gender. These comments total 1,512 tokens, and an average of 27.2 tokens per comment. This is a very small corpus, even by ‘specialised’ or ‘small’ corpus standards (Flowerdew, 2004; Koester, 2010). Nevertheless, patterns emerge from the data, and while these patterns cannot be generalised as being representative of all harassment experienced by children, they provide important and useful insights into this under-researched social problem.

**Method**

The free-text comments were extracted from the reports and saved as plain text (.txt) files. The corpus was part-of-speech (POS) tagged using TagAnt (Anthony, 2015) which uses the 58-tag Tree Tagger Tag set. Tagging a small corpus for part-of-speech is useful because although the frequencies of individual lexical items are likely to be low, POS tags occur with much higher frequencies (e.g. Poole, 2016: 581), and provide a richer dataset for analysis. Given that the focus of this paper is on the behaviours and actions reported in the free-text comments, the starting point for analysis is those words tagged as verbs. There are 24 different tags for verbs in TagAnt (all of which begin with ‘V’), and a wildcard query ‘V*’ was run in AntConc (Anthony, 2014) to identify all of the verbs in the corpus. There are 137 different verbs in the corpus, totalling 341 tokens. The 137 verbs were manually categorised according to different types of harassment behaviours to which they related. Like POS tagging, aggregating lexical items in semantic categories in this way allows us to identify frequency patterns that would not otherwise emerge (e.g. Baker, Gabrielatos and McEnery, 2013; Potts, Bednarek and Caple, 2015).

**Results**

Most of the verbs which appear in the data belong to one of two kinds. On the one hand, we find verbs which children use to describe the harassment they experienced, and on the other children are describing their own actions, both before and after the incident took place. By analysing these verbs separately, we gain an insight into the types of harassment directed towards children, what they were doing when they were harassed, and how they responded.

**Perpetrator’s actions**

In categorising the verbs that children attribute to others, six ‘types’ of harassment emerge from the reports (Table 1). The harassment type most commonly reported in the free-text comments are incidents involving people in cars beeping at the children, slowing down or stopping. These are closely followed by incidents with some verbal interaction between the harasser and the child. This often involves the child being called names and men initiating a dialogue with young girls, complimenting them and inviting them into their cars. Next are types of harassment that do not involve any verbal interaction, in which children are watched, stared at, and pictured/videoed by someone on a mobile phone, or smiled, waved or pointed at. In a small number of cases, children are followed, chased and even cornered by
<table>
<thead>
<tr>
<th>Harassment type</th>
<th>Freq.</th>
<th>Verbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Involving vehicle</td>
<td>29</td>
<td>beeped, stopped, slowed, turned (around), drove, speeding, honed, honked, driving, cycled, curbed.</td>
</tr>
<tr>
<td>Verbal interaction</td>
<td>25</td>
<td>said, called, shouted, laughing, asked, whispered, told, shouting, say, lafed, convincing, calling.</td>
</tr>
<tr>
<td>Being watched</td>
<td>15</td>
<td>looked, stared, watching, took, videoing, papped, staring.</td>
</tr>
<tr>
<td>Non-verbal interaction</td>
<td>12</td>
<td>smiled, waved, stuck, shrugged, showed, pulling (faces), pointing, pointed, bullied, whistled.</td>
</tr>
<tr>
<td>Being followed</td>
<td>9</td>
<td>followed, follow, following, cornering, coming, chased.</td>
</tr>
<tr>
<td>Physicality</td>
<td>5</td>
<td>grabbed, yanked, threw, hit.</td>
</tr>
</tbody>
</table>

Table 1. Categories of harassment expressed by verbs in the free-text reports

Children’s actions

Not all verbs in the reports are attributed to harassers; we can also learn about what children were doing before and in response to incidents (Table 2). By far the most common verb that children assign to themselves is walking (n=16). A concordance analysis of walking shows the situations in which children most frequently report experiencing harassment – when they were walking to and from school, home, their friends’ homes, or the shops (Figure 1).

<table>
<thead>
<tr>
<th>Children’s actions</th>
<th>Freq.</th>
<th>verbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before</td>
<td>28</td>
<td>walking, going, sitting, waiting, riding, left, leave, crossed, talking.</td>
</tr>
<tr>
<td>Response</td>
<td>26</td>
<td>ran, walked, moved, went, run, move, lost, cycling, covered, said, shouted, call, spoke, reported, rang.</td>
</tr>
</tbody>
</table>

Table 2. Children’s actions before the harassment incident and in response to it
Finally, the children also reported the ways in which they coped with or responded to harassment (Figure 2). Most often, they took passive or evasive action, including running, moving or walking away, which is line with existing research on women’s coping strategies for stranger harassment (Saunders et al., 2016). However, in addition to these passive strategies, children also report taking more active measures, including confronting their harassers, telling their parents and, in a very small number of cases, informing the police. These active strategies align with some of the response types less frequently employed by adults in response to harassment (e.g. Gruber, 1989; Magley, 2002).

Figure 1. Concordance lines for walking.

Figure 2. Concordance lines showing passive and active action taken by children.
Implications

A linguistic analysis of this pilot data gives us the first insights into street harassment of children in England, as we begin to have an understanding of how children perceive harassment, the different kinds of harassment they experience and how common they are, and how children cope with harassing behaviours from others. These findings have formed the basis of guidance documents for the support of young victims of street harassment prepared for the British Transport Police and Hollaback London, part of a global network dedicated to raising awareness of harassment and developing strategies to prevent it. This pilot also forms the foundation of a study with a much larger scope, and reinforces the opportunities afforded to researchers across disciplines by small, specialised corpora and corpus methods of linguistic analysis.

References


Ever since the ground-breaking work by John McH. Sinclair in the 1960s (Sinclair, 1966), collocation (the habitual co-occurrence of two words at a short distance from each other) has established itself as one of the mainstays of corpus-based research. Although J. R. Firth (1957/1968) is credited with making collocation “not just as an observable effect of language use, but as an important element of the causes of language patterns” (Barnbrook, Krishnamurthy, & Mason, 2013, p. 36), it is Sinclair who has been considered the ‘discoverer’ of collocation (Hoey, 2009, p. 39), due to the detailed evidence that his analyses of computer corpora provided in support of the existence of collocation in language use. A large number of studies have focused on collocation since then, looking at various aspects of the concept, from the association between lexical patterning and meaning (Moon & Sinclair, 1987), to the actual patterns of collocation for particular words (Stubbs, 2002), among other features. At the same time, research on collocation has generally ignored the relationship between collocation and text varieties (Biber, 2010, p. 245), focusing primarily on patterns of collocation that cut across register differences. An exception is Sinclair, Jones, and Daley (1970/2004), who compared collocations in a science magazine (New Scientist) and in conversation and found that the collocations could discriminate between the two registers. The authors concluded that:

from a linguistic point of view it is interesting to find that ‘strength of collocation’ provides a useful discriminant between different types of English and it would be interesting to see if the results were so encouraging for two texts which differ very little. (p. 133)

This paper attempts to fill the gap in collocation studies by reporting the results of a multidimensional study on collocations from a register perspective using the 450-million-word Corpus of Contemporary American English (Davies, 2012; see Table 1).

<table>
<thead>
<tr>
<th>Register</th>
<th>Tokens</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spoken</td>
<td>90,786,821</td>
<td>20.6%</td>
</tr>
<tr>
<td>Magazine</td>
<td>90,780,789</td>
<td>20.6%</td>
</tr>
<tr>
<td>Newspaper</td>
<td>87,131,579</td>
<td>19.8%</td>
</tr>
<tr>
<td>Academic</td>
<td>86,512,881</td>
<td>19.6%</td>
</tr>
<tr>
<td>Fiction</td>
<td>85,907,930</td>
<td>19.5%</td>
</tr>
<tr>
<td>Total</td>
<td>441,120,001</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 1: Composition of COCA 2012, downloadable full-text version

The method for this investigation was inspired by the multidimensional (MD) analysis of register variation, introduced by Biber in the 1980s (Biber, 1988) and subsequently developed by him and colleagues (cf. Berber Sardinha & Veirano Pinto, 2014). The goal of an MD analysis is to determine the dimensions or underlying parameters of variation in the data (see Friginal & Hardy, 2014, for an overview of the method). Several major differences exist between a mainstream MD analysis and the MD analysis carried out here. First, in this investigation, the units upon which the analysis was based were collocations and not texts—more specifically, pairs of words, with one representing a node and the other, a collocate (these nodes and collocates were selected from among the most frequent words in each register in COCA). Second, in this investigation, the measurements
taken for each unit were not text counts, but the log-dice, a word association statistic (Rychly, 2008) that gauged the attraction between the two words. Third, the factor scores were calculated for the collocates of each node rather than the texts in the corpus. Finally, the basis for the interpretation of the factors in this investigation was based primarily (but not solely) on lexical features revealed by their semantic preference (Stubbs, 2007), lexical sets (Sinclair & Jones, 1974/1996), word fields (Lehrer, 1974; Trier, 1931), ‘aboutness’ (Phillips, 1989; Scott, 2000; Yablo, 2016), topics (Berber Sardinha, 1997), and subject matter (Schütze, 1998), and not primarily on functional / communicative grounds.

A program designed for this project retrieved the most salient node-collocate pairs in each register. The resulting data matrix consisted of 3,511 columns (one for each node) and 23,602 rows (one for each collocate). Each cell in the data matrix contained the log-dice value for the node-collocate pair. The log-dice statistic measures the degree of lexical association between a node and a collocate in a span of four words on either side of the node. The dimensions were determined through a series of factor analyses carried out in SAS University Edition.

Nine dimensions of collocation were identified: 1. Literate discourse; 2. Oral discourse; 3. Objects, people, and actions; 4. Colloquial and informal language use; 5. Organizations and the government; 6. Politics and current affairs; 7. Feelings and emotions; 8. Cooking; and 9. Education. Some of the major collocations that typify each dimension are presented in Table 2.

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Literate discourse</td>
<td>[issue<del>n + relate-v], [factor</del>n + relate-v], [seem<del>v + appropriate-j], [specific-j + area</del>n], [assessment-n + tool~n]</td>
</tr>
<tr>
<td>2. Oral discourse</td>
<td>[want<del>v + know-v], [people</del>n + know-v], [want-v + say<del>v], [people</del>n + think-v], [think-v + go~v]</td>
</tr>
<tr>
<td>3. Objects, people, and actions</td>
<td>[stare-v + window<del>n], [stare-v + ceiling</del>n], [slide-v + open<del>j], [pull-v + trigger</del>n], [car~n + pull-v]</td>
</tr>
<tr>
<td>4. Colloquial and informal language use</td>
<td>[afraid-j + lose<del>v], [mama-n + papa</del>n], [mama-n + daddy<del>n], [mommy</del>n + daddy-n], [glad-j + hear~v]</td>
</tr>
<tr>
<td>5. Organizations and the government</td>
<td>[protection<del>n + agency-n], [official-n + say</del>v], [international-j + monetary<del>j], [national-j + association</del>n], [district-n + attorney~n]</td>
</tr>
<tr>
<td>6. Politics and current affairs</td>
<td>[other<del>j + politician-n], [decline</del>v + interview-v], [police<del>n + interview-v], [deserve</del>v + credit<del>n], [think</del>v + deserve~v]</td>
</tr>
<tr>
<td>7. Feelings and emotions</td>
<td>[feel<del>v + shame</del>n], [feel<del>v + guilt</del>n], [feel<del>v + rage</del>n], [face<del>n + rage</del>n], [feel~v + excitement-n]</td>
</tr>
<tr>
<td>8. Cooking</td>
<td>[mix-v + bowl<del>n], [mix-v + ingredient</del>n], [cup-n + sugar<del>n], [add-v + heat</del>n], [add-v + onion~n]</td>
</tr>
<tr>
<td>9. Education</td>
<td>[student~n + benefit-v], [rate-v +...]</td>
</tr>
</tbody>
</table>
Table 2: Typical collocations in each dimension (‘~’ indicates a node, and ‘-’ represents a collocate; order of node and collocates reflects order in which they are most often found in COCA).

| scale~n|, |expose-v + student~n|, |expose-v + child~n|, |educate-v + public~n| |

The greatest mean scores for the registers on each dimension were the following: in dimensions 1 and 9, ‘academic’; in dimension 2, ‘spoken’ and ‘fiction’; in dimensions 3, 4 and 7, ‘fiction’; in dimension 5, ‘newspaper’ and ‘academic’; in dimension 6, ‘spoken’; and in dimension 8, ‘magazines’. Despite these contrasts, the register differences were statistically negligible, due to the irregular distribution of collocation in language. This suggests that each individual dimension is not a reliable predictor of register differences. However, while collocations may not be strong predictor of register categories, it is possible that register categories are strong predictors of collocation. To this end, a discriminant function analysis (DFA) was employed, which used the factor scores of each collocate with each node on each dimension as input and produced discriminant equations that were used to place the collocation in its most likely register, based on its factor scores. A DFA was run in SPSS having as input the dimension scores of each collocation on each dimension as the dependent variable and the register categories as the independent variable. The ‘leave one out’ option was used in the DFA, and therefore each collocation was excluded from the model used to predict its classification, so as to prevent bias in the classification task. Four samples of different sizes were used—namely, 500, 1,000, 2,000, and 4,000 collocates per register; these consisted of the n collocates with the greatest scores per register. The best results were obtained with the 500-word-per-register sample, where the majority of collocations (56.7%) were assigned to their source registers at a rate nearly three times better than chance. The cases of cross-classification (where the collocations from one particular registers were attributed to a different register) were also examined. Magazine was the most cross-classified register, as its collocations were habitually assigned to spoken and newspaper. This suggests that magazine texts have both a ‘spoken-like’ and a ‘newspaper-like’ character. Cross-classification was not bi-directional, though. Newspaper collocations were less likely to be wrongly predicted as magazine collocations than the other way around. The least cross-classified register was academic, with 2/3 of its collocation being correctly attributed.

To summarize, this study presents a large-scale study of cross-register variation across American English collocations. Nine general groups of collocation (dimensions) were identified. When these nine dimensions are combined, they can predict the register from which a collocation occurs significantly better than chance. Overall, the results suggest that the use of collocation is sensitive to register constraints. The register differences associated with collocation use shown by this study provide another ‘nail in the coffin’ in the attempts to describe ‘general English’ or any other language, as if language were a homogenous whole.

References


In a language learning setting, concordancers are particularly useful for showing differences between patterns of different lexical items. Language learners may explore grammatical and lexical patterning in concordance lines or draw on results from measurements like collocation. One aspect of patterning which could be important but is hard for learners to explore, is the way similar lexical items may typically be used in rather different semantic contexts. Being able to recognise the typical semantic contexts of synonyms and words which have a common translation could be a useful step in distinguishing between words with a similar meaning. For words which seem to have more than one sense or more than one use in different domains, information about the patterning of the semantic contexts could also help language learners find distinctions. These kinds of patterning could also help language learners explore and uncover hidden connotation-like qualities which words may have resulting from common use with other words. An overview of the historical development of semantic prosody and related theories, including the influences of work by Sinclair, Louw, Stubbs, Hoey and others, is provided by Stewart (2010). Although there are some differences between conceptions and definitions of semantic prosody, semantic association and semantic preference (Hoey, 2005), they all include the possibility for emotive charging of words through their frequent use in specific contexts. In terms of the pedagogical implications of work in this area, Stewart argues there is a need for “serious improvements” in descriptive works for English language learning (2010, p. 263). The need to address the gap in English dictionary resources in China has also been highlighted as a high priority (Ping-Fang & Jing-Chun, 2009). In cross-linguistic work, it has been shown that the tendencies of semantic prosody of similar items across different languages can be quite different (Xiao & McEnery, 2006). Lack of access to information about typical uses of words and collocations and lack of access to information about tendencies of semantic prosody in dictionary resources and other descriptive works could account for some of the difficulties non-native speakers face in trying to produce language which matches the expectations and conventions of their intended audience.

With some assistance, learners can use concordancers to explore corpus examples with these questions in mind. When comparing words or phrases (particularly when the results can be viewed side-by-side on screen), some aspects of semantic association should be fairly clear. Language teachers would probably not want to specifically teach the linguistic terminology of semantic prosody, but visualizations such as collocation clouds often provide sufficient evidence of how some words tend to be associated with positive or negative collocates. More detailed analysis allows for narrower categorizations or groupings. Hunston introduces ways in which a hidden meaning of words may be deduced through analysis of concordance lines, an also provides examples of how different meanings of words can be observed by looking at patterns in the co-text (2002). When researchers are looking at data like these, there is a need for them to consider possible bias in the corpus because of the kinds of texts from which it was built, to consider “resonances of intertextuality” and to be aware that observational evidence
should usually be interpreted as “often” but “not always” (Hunston, 2007). When language learners are looking at data like these, they may need encouragement to open up to the possibility of looking for semantically-related items in the nearby context, and they may also need assistance in judging the strength of any attitudinal meaning. The question arises as to whether computational approaches could assist with this.

Within the field of Information Retrieval and Natural Language Processing, there are an increasing number of electronic resources for Sentiment Analysis holding information about associations between words and attitudes, emotions, etc. The General Inquirer lexicon, the Dictionary of Affect in Language, SentiWordNet, and WordNet-Affect are four widely used resources in this field (Devitt & Ahmad, 2013). Other resources include LIWC (Pennebaker, Chung, Ireland, Gonzales, & Booth, 2007) and USAS (Rayson, Archer, Piao, & McEnery, 2004). When making use of these resources, computational methods are often concerned with attempting to mark texts or text extracts as being associated with particular attitudes or used for searches within a specified semantic environment. Some applications of these resources are more related to information retrieval or monitoring of content; others provide additional tools for corpus linguistic analyses. When the object to be explored is a text (or a collection of texts), for example, WMatrix (Rayson, 2008) makes comparisons very clear by showing results of key semantic domains for one using the other as a reference corpus. However, when the object to be explored is a word, a collocation, or a pair of words with a similar meaning, semantic tags could also be used to provide summary information about tendencies the search query to be used in specific semantic contexts.

The Prime Machine (Jeaco, 2015) is a concordancer which was developed for students of English and their teachers as a resource for language learning. As well as search support features, it has several ways of encouraging learners to explore aspects of the environment of the lexical items they are interested in, and to compare these side by side with similar items. This paper reports on some further developments of The Prime Machine, which now draws on semantic tagging data to provide two kinds of information to learners regarding the neighbourhood of the words from their search query. They can then interact with these data in three ways: through tables or graphs showing proportions of concordance lines within positive or negative environments; and through filtering and comparing concordance lines.

The first way in which semantic tagging is used is by marking sentences in the corpus based on semantic tagging data from USAS (Rayson, et al., 2004). This is operationalized by counting only those semantic tags which meet a threshold of repetition within +/-1 sentence. In the concordancer, this wider context is easily accessible for each concordance line and it is also worth noting that while collocation information about words within windows of several words is certainly very important, Hoey has demonstrated how associations between words may occur (and be of importance) in wider contexts too (2014). Links based on thresholds of 2-8 repetitions are stored in the database, meaning that some fine-tuning is available. Since many items may have multiple semantic tags, this provides a straight-forward (albeit limited) means of "disambiguation", as only after a threshold number of items in the sentence have been found to share the same semantic tag will they be counted. The software uses log likelihood contingency
tables for each word and each collocation stored in the database (shown in Table 1), by creating a sub-corpus of sentences marked with each specific semantic tag. Figure 1 shows a screenshot of the cloud of results for *due to*.

**Table 1: Semantic Tag Contingency Table**

<table>
<thead>
<tr>
<th>Sub-corpus of sentences marked with a specific semantic tag</th>
<th>The rest of the corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>A = Count of word (or words in a collocation)</td>
<td>B = Total count of word (or words in a collocation) – A</td>
</tr>
<tr>
<td>C = Count of all words in sentences with the semantic tag</td>
<td>D = Total corpus size – C</td>
</tr>
</tbody>
</table>

![Figure 1: Screenshot of the Neighbourhood Cloud for *due to* in the BNC: Academic Sub-corpus.](image)

The second way semantic tagging is used is based on a narrower window (+/- 4 words), drawing on specific USAS tags associated with positivity or negativity and two lists of positive and negative words: 24 semantic tags from USAS (such as “A1.1.2 Damaging And Destroying” and “A1.4 Chance, Luck”), and the two other lists from as combined resources derived from the General Inquirer lexicon (GI, 2000), the NRC emotion lexicon (Mohammad & Turney, 2012) and a Loughran-McDonald wordlist (Bodnaruk, Loughran, & McDonald, 2015). Each word in the corpus is marked according to whether it occurs near these items. Since words on the list will always be marked, a further flag is used to indicate whether at least one other positive/negative item also occurs. The markers are then processed like the other features of lexical priming (as in earlier versions of the software): proportions are stored for each lexical item and the norms for the corpus, and a log likelihood measure is used to determine whether or not icons appear in the application to encourage users to explore this aspect. When looking up words in the corpus, if the positive or negative relationship is strong, the icon link used to explore this pops out on the icon bar. Figure 2 shows the graph of the results for *due to*, with the storm...
An icon at the bottom of the screen indicating it is strongly related to negative environments.

Figure 2: Screenshot of the Graph Showing the Proportion of Concordance Lines Marked in Positive and Negative Environments for *due to* in the BNC: Academic Sub-corpus.

It is argued that this approach provides a means of helping language learners explore pertinent features of the neighbourhood in the concordance lines, giving a helping hand for their concordance line analysis and ultimately for their insights and awareness of patterns of language use. Details about access to the software will be available from [www.theprimemachine.com](http://www.theprimemachine.com).
References


From corpora to learners' dictionaries:
A case study of the function word otherwise
Stephen Coffey (University of Pisa, Italy)

1. Introduction

In this study I examine descriptions of the function word otherwise in five corpus-based learners' dictionaries, and compare these descriptions with findings of an analysis carried out on a sample of contexts for otherwise in the British National Corpus (BNC).

The dictionaries examined are: Cambridge Advanced Learner's Dictionary (cd-rom, hereafter CALD4), Collins Cobuild Advanced Dictionary of English (cd-rom, COB5), Longman Dictionary of Contemporary English (online, LDOCE6), Macmillan English Dictionary for Advanced Learners (online, MEDAL), and Oxford Advanced Learner's Dictionary (cd-rom, OALD9). These are broadly comparable dictionaries, all being produced by UK-based publishers and marketed with higher-level learners in mind.

The BNC was searched using the BNCweb interface (http://bncweb.lancs.ac.uk). The corpus contains 8622 tokens of otherwise, and a 5% random sample (431 tokens) was analyzed by reading through extended contexts for each token.

2. Dictionary analysis of otherwise: recognized senses and functions

The five dictionaries differ as regards the number of senses/functions assigned to the word otherwise. Ignoring those presented in terms of fixed phraseology (e.g. or otherwise, as in 'accidentally or otherwise'), there are either three, four or five senses (see Table 1). Grammatical labelling also differs.

<table>
<thead>
<tr>
<th>SENSES</th>
<th>GRAMMATICAL LABEL/S</th>
</tr>
</thead>
<tbody>
<tr>
<td>CALD4</td>
<td>4conj / adv / adv / adj after verb</td>
</tr>
<tr>
<td>COB5</td>
<td>4adv with clause</td>
</tr>
<tr>
<td>LDOCE6</td>
<td>5sentence adv</td>
</tr>
<tr>
<td>MEDAL</td>
<td>51 general heading: adverb (with a note about use as adv or sent. adv)</td>
</tr>
<tr>
<td>OALD9</td>
<td>31 general heading: adverb</td>
</tr>
</tbody>
</table>

Table 1. Otherwise in dictionaries: number of sense divisions, and grammatical labels

The various explanations of otherwise in the dictionaries are now discussed. In order to facilitate discussion, they are presented in terms of five specific senses/functions, being those which have most clearly emerged from a comparative analysis of the relative dictionary entries.

1) 'Except for what has just been mentioned'
This meaning (the wording comes from LDOCE6) is common to all dictionaries. Furthermore, four dictionaries include examples of otherwise being used in the following two, grammatically different, ways (the examples are from CALD4): The bike needs a new saddle, but otherwise it's in good condition. ♦ The poor sound quality ruined an otherwise splendid
It is to be noted, however, that only LDOCE6 explicitly separates the two grammatical uses, labelling the first as a 'sentence adverb' and the second as the pattern 'adverb + adjective/adverb'.

2) 'Other ways of doing something'
Three dictionaries (COB5, LDOCE6, MEDAL) include a usage described as follows in COB5: 'You use otherwise to indicate that other ways of doing something are possible in addition to the way already mentioned: The studio could punish its players by keeping them out of work, and otherwise controlling their lives.' CALD4 and OALD9 also include examples of this type, but they are within more general sub-entries of adverbial usage.

3) 'In a different or opposite way from what has been mentioned'
This meaning (the wording is from MEDAL) is presented in COB5, LDOCE6, and MEDAL. An example from LDOCE6 is: The government claims that the economy is improving, but this survey suggests otherwise. COB5 associates this usage with written English, and LDOCE6 presents it (phraseologically) as 'say/think/decide etc otherwise'. CALD4 and OALD9 also include examples which correspond to this usage, but the examples are within more general sub-entries.

4) Otherwise as a predicative adjective
One dictionary (CALD4) includes a usage in which otherwise is 'used to show that something is completely different from what you think it is or from what was previously stated: He might have told you he was a qualified electrician, but the truth is quite otherwise.' Meaningwise, this usage is similar to 3).

5) Conditional otherwise: 'If not'
All dictionaries include a use which has the word if in the relative explanation/s, though they differ considerably as to how this general function is broken down into more specific uses. Time reference is one variable. For example, future and present reference can both be seen in the explanation and examples in OALD9: 'used to state what the result would be if something did not happen or if the situation were different: Shut the window, otherwise it'll get too cold in here. ♦ We're committed to the project. We wouldn't be here otherwise. LDOCE6, by contrast, distinguishes between future and past reference, an example of the latter being, We were delayed at the airport. Otherwise we would have been here by lunch time. The explanation in CALD4 refers only to the future, but introduces a pragmatically more specific description: 'used after an order or suggestion to show what the result will be if you do not follow that order or suggestion'; one of the examples is: Phone home, otherwise your parents will start to worry.

Still with 'conditional otherwise', in two dictionaries (LDOCE6 and MEDAL), the idea of negative consequence is introduced: in LDOCE6, for example, we read that otherwise can be 'used when saying what bad thing will happen if something is not done'. Lastly, MEDAL introduces the notion of necessary consequence: 'used when you are trying to show that something must be true, by saying that the situation would be different if it was not true', Of course they're interested. Otherwise they wouldn't be asking about prices. ♦ He must be fairly intelligent, otherwise he wouldn't have got into university.

---

1 Examples in corpus-informed dictionaries are not necessarily taken directly from corpora; most dictionaries adapt many citations in order to make them clearer to language learners. In the case of a word like otherwise, several lines of authentic text may be necessary in order to clarify meaning.
3. BNC-derived data and dictionary data

The various functions of otherwise listed above were confirmed by BNC data, but other aspects of usage were noted which are absent from dictionaries or not explicitly pointed out; I will give a few examples in relation to general usages 2) and 5) above. To start with the former, in this usage otherwise is only found in the dictionaries with verbs, whereas in the corpus sample it is also found with adjectives or adjectival phrases (e.g. '... may prove to be untraceable or otherwise impracticable to pursue.' Also, the sequences 'and otherwise' and 'or otherwise' appear to be an integral part of this sense, but this phraseology (with regard to this sense) is not highlighted in dictionaries.

Turning now to the set of uses grouped above under the broad label 'conditional otherwise', in the dictionary examples otherwise appears 12 out of 15 times in clause-initial position and twice in clause-final position. This is a useful dictionary strategy which enables two simple clause meanings to be clearly contrasted and linked by the word otherwise. In the corpus sample, however, otherwise is also frequently found within verb phrases (e.g. '... who would otherwise have died'), and sometimes after conjunctions (e.g. 'because otherwise ...'). Furthermore, and in relation to written texts, there is no comment in dictionaries on the fact that when otherwise is clause-initial, it can either follow a full-stop (or semi-colon) and thus introduce a sentence-like unit, or else follow a comma, thus introducing a clause which is itself part of a sentence.

4. Discussion and suggestions

Otherwise is well distributed throughout modern British English, at least judging from frequency figures in the BNC. Leech et al (2001, p. 84) list it as having a frequency of 88 per million words, both overall and individually for the written and spoken sub-corpora; it also appears in the full (100/100) range of 'sectors' of the BNC and with a high dispersion value (95). It is interesting, therefore, that the corpus-based dictionaries examined in this study do not come to closer agreement on the description of otherwise, and also that there are aspects of its usage which are not included in any of them. The probable reason for this is that only a limited amount of space is devoted to this particular lexical item, and choices have therefore been made as to what data to include. However, while this may be a valid reason where print dictionaries are concerned, it is more difficult to justify elsewhere. Otherwise is a multi-faceted word, and deserves a richer description. It might also benefit from being recognized as a 'one-member word class' (Sinclair, 1999, p. 165). Otherwise, it is just another 'adverb'.

References

Candidate Knowledge? Exploring epistemic claims in scientific writing: A corpus-driven approach
Garry Plappert (University of Birmingham, United Kingdom)

Introduction

Whilst the identification of hedging devices has proven to be a very useful and successful enterprise within applied linguistics, it has been argued that the study of these devices has become concentrated onto a small group of the ‘usual suspects’ (Groom, 2007; 2010; Plappert, 2012) of words and structures that are known to have an epistemic effect in a claim or proposition. As such, linguistic markers of modality such as modal verbs (eg: may, might, can, could), modal adjectives (eg: possibly, probably) and n-grams identified as functioning as hedges (such as it is possible that and it is likely that) often form the starting place for analysis of the linguistic aspects of epistemology. This impasse has been compounded by a plethora of corpus-based studies (eg: Hunston, 1995; Noguchi et al., 2006; Thompson and Ye, 1996; Williams, 1996; Chi-Hua, 1999 and cf. Hyland 1998), which, whilst providing excellent empirically based descriptions of known epistemic structures, are unlikely to contribute to the discovery of additional or unknown epistemic devices.

In this paper I will argue, in agreement with Groom (2007; 2010), that the answer to this impasse is to explore corpus-driven methods of analysis in order to uncover new or unexpected epistemic devices in English. Through an inductive analysis of four clusters, I demonstrate that it is possible to discover a number of additional strategies for nuancing claims, which are not typically mentioned in seemingly exhaustive studies such as Hyland (1998). I also argue that the peripheral presence of the ‘usual suspects’ in the context of nodes such as tumor suppressor gene, mutations in the gene encoding and loss-of-function mutations raises the possibility that the epistemic devices of which we are already aware may be far more marginal phenomena than we currently assume.

Background

The study of epistemology within Applied Linguistics has focused on the linguistic devices used to mitigate claims (cf. Hyland 1998), though the term used for this phenomenon has varied considerably. Thus Hyland (1998) is able to identify studies of hedges (Lakoff, 1972) as well as ‘compromisers (James, 1983), downtoners (Quirk et al, 1972), weakeners (Brown & Levinson, 1987), downgraders (House & Kasper, 1981), softeners (Crystal & Davy, 1975), backgrounding terms (Low, 1996) and pragmatic devices (Stubbe & Holmes, 1995)’ (1998:9, my italics) as constituting what he wishes to call hedging. This subject, then, has undoubtedly received plentiful coverage in Applied Linguistics and work focused on identifying or analyzing hedging in academic discourse has become so common that Groom (2007)
has identified (rather desparingly) the ‘usual suspects’ of corpus study on this subject: ‘A glance at the recent literature identifies report clauses and other attributive forms [...] modal verbs and other hedging devices [...] and extraposed complement clauses and other kinds of that-clause [...] as being amongst the usual suspects’ (2007: 40). The advantage of this approach for the large scale analysis of written academic discourse is that the seemingly exhaustive lists of hedging devices provided by works such as Hyland (1998) and (2009) provide a clear and labour-saving basis for selecting and analysing items from wordlists, allowing the analyst to proceed with collocation or concordance line based description. However, such studies of known hedging devices are by their very nature unlikely to widen or extend the very list from which they are chosen: the list of known hedging devices.

Methods

The leading journal in the field of genetics is *Nature Genetics* (29.648 Thomson Reuters 2014, accessed February 2014) and it was decided that the corpus for this study would be comprised of texts from this journal. The texts for this study came from a ten-year period (1999-2008 inclusive) and were collected together in a corpus known as genecorp. In total genecorp contains 2,979 texts from the journal *Nature Genetics*, spanning from 1999-2008.

In order to carry out a ‘bottom-up’ analysis of claims made in genecorp the following procedure was adopted:

1. Generation of keywords using *BNC World* as reference corpus
2. Generation of clusters containing the ten most key keywords
3. Selection of all clusters containing three lexical items from (2)
4. Collocation analysis of tri-lexical clusters from (3)
5. Concordance line analysis of tri-lexical clusters from (3)
6. Form generalisations about geneticists epistemic practices based on the evidence of (4) and (5)
7. Inspect whole corpus frequencies where possible to check the plausibility of (6)

Results

The following table exemplifies the results of this study by summarising the verb phrase patterns found in relation to node phrases containing mutations and attempts to describe the epistemic function of these:

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Epistemic Function</th>
<th>Examples of forms identified</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAUSE group</td>
<td>To make a causal claim involving mutations</td>
<td>CAUSE*; LEAD* to; IMPAIR*; are due to; PRODUCE*;RESULT* + in; RESULT* + from; STOP*; TRIGGER*; UNDERLIE*</td>
</tr>
</tbody>
</table>
PREDISPOSE group | To posit a causative connection between mutations and a disorder that falls short of a full causative claim | PREDISPOSE*; INVOLVE*  
ASSOCIATED group | To express an association between mutations and a disorder without expressing a causal connection | ASSOCIATE*; LINK*;  
COPULA group | To identify mutations | is; are  
IMPLICATURE group | To juxtapose mutations with a disorder without characterizing the connection between the two linguistically | have; has  
EFFECTS and CONSEQUENCES group | To discuss the effects of mutations; To assess the effects of mutations; To speculate as to the effects of mutations | EFFECT*; CONSEQUENCE*  

**Figure 1: Summary of results of corpus-driven analysis for loss of function mutations and mutations in the gene encoding**

The inductive analysis of the node phrases in this article demonstrates a range of different epistemic claims. The prevalence of unhedged claims was clear in all nodes discussed but when geneticists sought to limit their claims they only rarely used the ‘usual suspects’ to do so. Rather, when they did seek to nuance claims they tended to use either much less familiar explicit hedging devices (such as putative or candidate) or otherwise typically modified the verbal group used to form the claim, leading to claims such as is linked to or is associated with instead of is caused by or other unhedged causal claims. These possibilities threaten the usefulness of general academic wordlists proposed in works such as Coxhead (2000); Simpson-Vlach and Ellis (2010) and Gardner and Davies (2014); and lend support to previous critiques of such lists (cf. Hyland and Tse 2007) which have argued that considerable disciplinary variation is being glossed over in the attempt to produce a universally usefully ‘general’ list of academic words or structures.

**References**


Building a Polish Corpus of Synesthetic Metaphors
Magdalena Zawisławska (University of Warsaw, Poland)

Introduction

The research project Synamet described in this paper aims at creating a semantically and grammatically annotated corpus of Polish synesthetic metaphors. The paper outlines the procedure followed in the project, the main problems with the metaphor annotation, and the preliminary results.

Synesthetic metaphor is when a perceptual sensation (e.g. olfactory) is described by lexemes that primarily activate another sense (e.g. taste), cf. sweet aroma. A broader definition includes the mapping of perception experience on more abstract domains, cf. sweet love. Synesthetic metaphors can provide valuable material for preliminary research because their common use in language guarantees that the material collected will be sufficiently rich and varied (which should help later to broaden the analysis to include other types of metaphor). In addition, restricting the research field to the domain of sensory perception will ensure the precise indication of the research scope.

The most important outcomes of the Synamet project are: creation of a valuable source of linguistic data, evaluation of existing models of metaphor and models of synesthesia, and so an important contribution to the redefinition of figurative language.

Method

Most recent corpus-based studies utilize the CMT, formulated by G. Lakoff and M. Johnson (1988). The authors view metaphor as a primarily conceptual phenomenon consisting of mapping across domains (from the source domain onto the target domain). Recently, in some works on metaphor, a ‘domain’ has been replaced with the concept of frame (Dancygier and Sweetser 2014). Likewise, in the project MetaNet: A Multilingual Metaphor Repository two different methodologies are employed: the CMT as well as the FrameNet ontology.

Fillmore posits that the meaning of lexical units, phrases, grammatical and syntactic constructions resides in frames—schematic phenomena such as our beliefs, experiences or typical actions (Fillmore 1982). The metaphorization process can be described as frame shifting. Coulson (2001) defines this phenomenon as a “semantic reanalysis process that reorganizes existing information into a new frame.” It means that some elements of a frame evoking specific sensations (e.g. smell) may become reorganized under the influence of a lexeme activating a frame of some other sensory perception (e.g. hearing). See, for example:

*Dochodzi zapach delikatnych kwiatów, szyprowy oddech mchu, a baza mruczy rozkosznie delikatnym piżmem.*

‘The smell of delicate flowers is drawing near, the chypre breath of moss, and the base note purrs contentedly with delicate musk.’
In the Synamet project, frames have been built up from scratch. The existing FrameNets are not suited to this purpose as there are too many semantic and grammatical differences between Polish and English. In the Synamet corpus, the frames and their elements are adjusted to the texts analyzed—that is, the frame coordinator adds new frames or their elements when the annotators signal that such modification is needed. At present, the corpus tool ATOS contains 6 perceptual frames (SIGHT, HEARING, TOUCH, SMELL, TASTE and MULTIMODAL PERCEPTION—for a sensation that activates several senses, e.g. weight or consistency), and 55 non-perceptual frames (e.g. MAN, ARCHITECTURE, WILD ANIMAL, DOMESTIC ANIMAL, PLANT, SPACE, TIME, ART, SOCIETY, ARMY, HAZARD, etc.). The frames in the project are understood as an analytical tool, not as real conceptual knowledge units, and as they are derived from linguistics data, they are not universal. Every frame element is marked with a typical lexical example, e.g. MAN/emotion (anger), WILD ANIMAL/part of animal (claw).

**Material used in the corpus and the tool for analysis**

According to Wering et al. (2006), a metaphor is synesthetic only when its source domain pertains to perception (visual, auditory, olfactory, tactile, or gustatory). If the target domain does not evoke perception, we can talk of a *weak synesthetic metaphor*. If both the source and the target domain evoke perception, we are dealing with a *strong synesthetic metaphor* (Werning, Fleischhauer, Beşeoğlu 2006). In Synamet both types of metaphors are annotated.

Since metaphors basically work within a context, the annotators analyze whole texts. The texts are excerpted from blogs devoted to perfume (SMELL), wine, beer, cigars, Yerba Mate, tea or coffee (TASTE, SMELL, SIGHT), as well as culinary blogs (TASTE, SIGHT), music blogs (HEARING), art blogs (SIGHT), massage and wellness blogs (TOUCH).

For the analysis, a dedicated tool called ATOS is used (Annotation Tool of Synesthetic Metaphor). The procedure of annotation includes:

1) Extraction of a metaphorical unit, e.g.: *tannin smooth*.
2) Correction of the text phrase (if needed), e.g.: *tannin is smooth*.
3) Definition of the referent, e.g.: taste (of a wine).
4) Description of the phrase type: nominal predicate.
5) Selection of the metaphor type: strong (because both frames are perceptual, (see Werning et al. 2006).
6) Selection of metaphor category: simple synesthesia.
7) Definition of the semantic head of the phrase: *tannin*.
8) Description of the source frame: TOUCH.
9) Selection of the source frame element: TEXTURE.
10) Description of the source frame evoking expression: *smooth*, part of speech: adjective.
11) Description of the target frame: TASTE.
12) Selection of the target frame element: TASTE COMPONENT.
13) Description of the target frame evoking expression: *tannin*, part of speech: noun.
Preliminary results of annotation

The statistics produced for 2915 metaphorical units (42% of the planned corpus) show some interesting features of synesthetic metaphors in Polish. The most frequent source perceptual frames are SIGHT (445 units) and TOUCH (297 units). The result is quite concerning, because in all models of synesthesia (Ullman 1962, Judycka 1963, Williams 1976, Classens 1993, Rogowska 2007) the most basic sense is touch.

Table 1. Source perceptual frames

<table>
<thead>
<tr>
<th>FRAME</th>
<th>UNITS</th>
</tr>
</thead>
<tbody>
<tr>
<td>WZROK (SIGHT)</td>
<td>445</td>
</tr>
<tr>
<td>DOTYK (TOUCH)</td>
<td>297</td>
</tr>
<tr>
<td>SŁUCH (HEARING)</td>
<td>242</td>
</tr>
<tr>
<td>PERCEPCJA ZŁOŻONA (MULTIMODAL PERCEPTION)</td>
<td>227</td>
</tr>
<tr>
<td>SMAK (TASTE)</td>
<td>181</td>
</tr>
<tr>
<td>ZAPACH (SMELL)</td>
<td>12</td>
</tr>
</tbody>
</table>

Less surprising are the most frequent target perceptual frames: in this case, the ultimate recipient is SMELL (1519). From the diachronic perspective, in Polish all olfactory expressions are metaphorical (see Judycka 1963).

Table 2. Target perceptual frames

<table>
<thead>
<tr>
<th>FRAME</th>
<th>UNITS</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZAPACH (SMELL)</td>
<td>1519</td>
</tr>
<tr>
<td>SŁUCH (HEARING)</td>
<td>524</td>
</tr>
<tr>
<td>SMAK (TASTE)</td>
<td>489</td>
</tr>
<tr>
<td>WZROK (SIGHT)</td>
<td>72</td>
</tr>
<tr>
<td>DOTYK (TOUCH)</td>
<td>10</td>
</tr>
<tr>
<td>PERCEPCJA ZŁOŻONA (MULTIMODAL PERCEPTION)</td>
<td>5</td>
</tr>
</tbody>
</table>

The most typical pairs of perceptual frames are: HEARING -> SMELL (203), SIGHT -> SMELL (184), TOUCH -> SMELL (155) and TASTE -> SMELL (151). These results are also contrary to the existing models of synesthesia. Judycka (1963) claims that synesthetic metaphors involving hearing and smell are very rare in Polish. She also contends that primarily visual expressions are used mostly for description of various sound features (e.g. purity, spaciousness of a tone).

Table 3. Pairs of source and target perceptual frames

<table>
<thead>
<tr>
<th>SOURCE FRAME</th>
<th>TARGET FRAME</th>
<th>UNITS</th>
</tr>
</thead>
<tbody>
<tr>
<td>SŁUCH (HEARING)</td>
<td>ZAPACH (SMELL)</td>
<td>203</td>
</tr>
<tr>
<td>WZROK (SIGHT)</td>
<td>ZAPACH (SMELL)</td>
<td>184</td>
</tr>
<tr>
<td>DOTYK (TOUCH)</td>
<td>ZAPACH (SMELL)</td>
<td>155</td>
</tr>
<tr>
<td>SMAK (TASTE)</td>
<td>ZAPACH (SMELL)</td>
<td>151</td>
</tr>
<tr>
<td>WZROK (SIGHT)</td>
<td>SŁUCH (HEARING)</td>
<td>97</td>
</tr>
<tr>
<td>PERCEPCJA ZŁOŻONA (MULTIMODAL PERCEPTION)</td>
<td>ZAPACH (SMELL)</td>
<td>85</td>
</tr>
<tr>
<td>WZROK (SIGHT)</td>
<td>SMAK (TASTE)</td>
<td>75</td>
</tr>
</tbody>
</table>
In the case of non-perceptual source frames, the most common are personification and ontological metaphors. There are also quite frequent metaphors with the following target frames: CLOTHES, ARCHITECTURE, SPACE and PLANT. The unexpected, but fairly frequent source frames, are those which are very abstract, such as PHYSICAL QUANTITIES, ABSTRACT IDEAS or LANGUAGE.

Table 4. Source non-perceptual frames

<table>
<thead>
<tr>
<th>FRAME</th>
<th>UNITS</th>
</tr>
</thead>
<tbody>
<tr>
<td>CZŁOWIEK (MAN)</td>
<td>557</td>
</tr>
<tr>
<td>RZECZ (THING)</td>
<td>137</td>
</tr>
<tr>
<td>UBRANIE (CLOTHES)</td>
<td>68</td>
</tr>
<tr>
<td>ARCHITEKTURA (ARCHITECTURE)</td>
<td>62</td>
</tr>
<tr>
<td>PRZESTRZEŃ (SPACE)</td>
<td>56</td>
</tr>
<tr>
<td>ROŚLINA (PLANT)</td>
<td>52</td>
</tr>
<tr>
<td>SZTUKA (ART)</td>
<td>38</td>
</tr>
<tr>
<td>SPOŁECZENSTWO (SOCIETY)</td>
<td>37</td>
</tr>
<tr>
<td>ZWIERZĘ DZIKIE (WILD ANIMAL)</td>
<td>31</td>
</tr>
<tr>
<td>WIELKOŚCI FIZYCZNE (PHYSICAL QUANTITIES)</td>
<td>30</td>
</tr>
<tr>
<td>POGODA (WEATHER)</td>
<td>29</td>
</tr>
<tr>
<td>ŻYWIOŁY (ELEMENTS)</td>
<td>29</td>
</tr>
<tr>
<td>ZDARZENIE/AKCJA (EVENT/ACTION)</td>
<td>21</td>
</tr>
<tr>
<td>POJĘCIA ABSTRAKCYJNE (ABSTRACT IDEAS)</td>
<td>20</td>
</tr>
<tr>
<td>WOJSKO (ARMY)</td>
<td>19</td>
</tr>
<tr>
<td>DOM (HOME)</td>
<td>18</td>
</tr>
<tr>
<td>PODRÓŻ (TRAVEL)</td>
<td>16</td>
</tr>
<tr>
<td>AKWEN (BASIN)</td>
<td>15</td>
</tr>
<tr>
<td>JĘZYK (LANGUAGE)</td>
<td>15</td>
</tr>
<tr>
<td>ZWIERZĘ DOMOWE (DOMESTIC ANIMAL)</td>
<td>15</td>
</tr>
</tbody>
</table>

The pairs of non-perceptual source frame and perceptual target frames show that the most personified senses are SMELL (283), TASTE (169) and HEARING (88). In turn, ontological metaphors are typical for HEARING (88) and SMELL (58).

Table 5. Pairs of non-perceptual and perceptual frames

<table>
<thead>
<tr>
<th>SOURCE FRAME</th>
<th>TARGET FRAME</th>
<th>UNITS</th>
</tr>
</thead>
<tbody>
<tr>
<td>CZŁOWIEK (MAN)</td>
<td>ZAPACH (SMELL)</td>
<td>283</td>
</tr>
<tr>
<td>CZŁOWIEK (MAN)</td>
<td>SMAK (TASTE)</td>
<td>169</td>
</tr>
<tr>
<td>CZŁOWIEK (MAN)</td>
<td>SŁUCH (HEARING)</td>
<td>88</td>
</tr>
<tr>
<td>RZECZ (THING)</td>
<td>SŁUCH (HEARING)</td>
<td>58</td>
</tr>
</tbody>
</table>
The most frequent source frame elements used in the synesthetic metaphors analyzed are: type of taste (TASTE FRAME), e.g. sweet, bitter (211), element of tone (HEARING FRAME), e.g. note (205), weight (MULTIMODAL PERCEPTION FRAME), e.g. light, heavy (196), temperature (TOUCH FRAME), e.g. cold, warm (173), part of the body (MAN FRAME), e.g. heart, head (127).

Acknowledgments

This paper is funded by the National Science Centre in Poland under the project no. 2014/15/B/HS2/00182 titled: Synamet – the Microcorpus of Synaesthetic Metaphors. Towards a Formal Description and Efficient Methods of Analysis of Metaphors in Discourse.

References

The discursive representation of the US same-sex ruling:  
A corpus-based investigation of news values  
in the US, UK and Italian press  

Marco Venuti (University of Catania, Italy) and Antonio Fruttaldo (University of Naples Federico II, Italy)

On June 26, 2015, the US Supreme Court ruled on the Obergefell v. Hodges case that led to the recognition of same-sex marriage in all fifty States, declaring it a constitutional right under the Fourteenth Amendment. The event received massive media coverage and soon became a major topic of animated discussions on digital media platforms.

The media coverage and its worldwide resonance resulted in a(n involuntary) deviance amplification effect (Cohen 2002), framing the event in such a way that moral panic was an inevitable consequence (Hall et al. 1978; McEnery 2006).

Drawing on these observations, our investigation focuses on the cross-cultural discursive representation of key actors and events concerning the US Supreme Court ruling in leading US, UK and Italian newspapers. In particular, our contribution is based on the analysis of all the articles published in the timespan that goes from June 26, 2015 to July 3, 2015 by the following newspapers:

- UK corpus (UK_NewsMar, 33,717 tokens [42 articles]): The Guardian, The Daily Telegraph, and The Times;
- Italian corpus (ITA_NewsMar, 22,633 tokens [43 articles]): la Repubblica, Corriere della Sera, Il Messaggero, il Giornale, il Fatto Quotidiano, and Libero.

The selection was made both on the basis of the printed and online circulation of each newspaper in the respective countries, and their widespread use of digital media (see the next paragraph on this aspect of our research).

The data were collected by using the online database LexisNexis. However, due to the unavailability on LexisNexis of the sources previously mentioned for the Italian component of the NewsMar, Google News was used in order to collect these data. BootCat (Baroni and Bernardini 2004) was then used so as to collect automatically the news stories displayed thanks to Google News.

The corpora thus collected were then uploaded to the online corpus analysis platform Sketch Engine (Kilgarriff et al. 2004; Kilgarriff et al. 2014) in order to analyse the discursive construction of the event in relation to news framing.

Our investigation is part of a wider independent research project based on the cross-cultural analysis of news values in the press. News values have been traditionally described as “the factors that take an event into the news” (Bednarek and Caple 2012a: 39), that is, in terms of the factors that make a news story newsworthy. Thus, they have been considered as linked to news selection, since as Bell (1991) argues, they are “values by which one ‘fact’ is judged more newsworthy than another” (Bell 1991: 155). In the Journalism and Communication literature, this approach to news values is confirmed, since they are “typically defined as properties of events or stories or as criteria/principles that are applied by news workers in order to select events or stories as news or to choose the structure and order of reporting” (Bednarek and Caple 2014: 2).
While news values have generally been approached from the point of view of their material (i.e., as properties of the event in itself; see Galtung and Ruge 1965) or their cognitive construction (van Dijk 1988), Bednarek and Caple’s (2012a) view is mainly linked to their Discursive News Values Analysis (DNVA) approach (Bednarek and Caple 2017), which investigates “how newsworthiness is construed and established through discourse” (Bednarek and Caple 2012b: 104). This is due to the fact that, as Vasterman (1995) argues:

[...] news is not out there, journalists do not report news, they produce news. They construct it, they construct facts, they construct statements and they construct a context in which these facts make sense. They reconstruct ‘a’ reality.

Thus, a discursive perspective on news values can allow us to “systematically investigate how these values are constructed in the different types of textual material involved in the news process” (Bednarek and Caple 2012b: 104). This approach allows the authors to highlight given textual traces that can let us see how news values are realised in the news discourse.

From this observation, we must also underline that, as Bell (1991) highlights, these are values and, as such, “[t]hey are not neutral, but reflect ideologies and priorities held in society” (Bell 1991: 156, italics in the original). Indeed, Cotter (2010) refers to them as “ideological factors” (Cotter 2010: 8, 67), since they can be used to both reinforce “an ‘ideology’ about what counts as news” (Cotter 2010: 67) and to strengthen “other ideologies (rather than just an ideology of what is newsworthy)” (Bednarek and Caple 2014: 3, italics in the original). However, while scholars working within Critical Discourse Analysis have repeatedly hinted at their ideological nature (van Dijk 1988; Fowler 1991; Richardson 2007; Machin and Mayr 2012), as Bednarek and Caple (2014) highlight, news values have not been comprehensively researched in this field of investigation.

Thus, the aim of our research project is to demonstrate how given news values (Bell 1991; Bednarek and Caple 2012a, 2012b, 2014, 2017) are disseminated and/or ‘picked up’ by media institutions in their professional experience of what counts as news. Indeed, analysing if and how journalists reinforce and reinterpret these ideologies may help researchers “find out what values are emphasised (foregrounded), rare or absent (backgrounded)” (Bednarek and Caple 2014: 6).

In order to achieve this, corpus linguistic methodologies can help us when approaching a large amount of data, as Bednarek and Caple (2014) and Potts, Bednarek and Caple (2015) point out. Indeed, through the use of corpus linguistic methodologies, we can gain “first insights into a conventionalised repertoire of rhetoric of newsworthiness” (Bednarek and Caple 2014: 14) in the case of corpora representative of specific media events. Thus, if “every journalist and every editor will have a different interpretation of what is newsworthy” (Rau 2010: 15), corpus linguistic techniques can help researchers identify “what kind of discursive devices are repeatedly used […] to construct different news values” (Bednarek and Caple 2014: 16) and, consequently, they can take us to the backstage of the news production process. In this way, the combination of Discursive News Values Analysis (DNVA; Bednarek 2016a, 2016b; Bednarek and Caple 2017) and corpus linguistic methodologies can be used to better define how news stories are reported since, by underlining what is newsworthy for a particular news organisation, they can help researchers ‘sneak a peek’ into the professional practices at the very heart of the news production process.

In order to carry out our analysis of the news values that are routinely enhanced in our corpora, we have resorted to the Word list tool available on the Sketch Engine online
platform. A cut-off point of minimum frequency of five occurrences has been imposed, since as previously argued our investigation wants to highlight discursive patterns that are routinely used by the media organisations under investigation.

In order to further ensure this, that is, that the selection of given items displayed in the Word list was not only due to their frequency but also to their dispersion in the corpora under investigation, we have decided to make use of the Average Reduced Frequency (ARF; Savický and Hlaváčová 2002). This is something different from the methodology adopted by Bednarek and Caple (2014) and Potts, Bednarek and Caple (2015), since they only take under consideration as a parameter in the analysis of news values the raw frequency of given words in the corpora they set out to investigate, without considering the dispersion of the phenomena highlighted.

The word lists thus computed were not, however, contrasted with a reference corpus (i.e., no keyword analysis was performed), since the scope of our analysis was not to contrastively highlight differences among genres in news discourse but, rather, see how given items in our corpora were explicitly and routinely used in order to construe the newsworthiness surrounding the US Supreme Court ruling. Indeed, a comparison with other corpora might have altered the results, due to the nature of the genres contrasted per se and not to the way newsworthiness is construed in the US_NewsMar, UK_NewsMar, and ITA_NewsMar.

In order to assign a news value to given items in the word lists computed according to the methodology previously described, a collocation analysis was also performed for each of these elements, so as to explore the semantic prosodies/preferences (Louw 1993; Stubbs 2002; Partington 2004) of given phenomena and attribute a specific news value.

In a preliminary study, focusing on a qualitative analysis of the first news stories on the US Supreme Court ruling published by the newspapers under investigation on June 26, 2015, our findings have identified how different cultures tend to highlight different sets of news values in the three countries and different newspapers under investigation. More specifically, this preliminary study has highlighted that, while the Italian newspapers tend to represent the event in terms of the news values of Eliteness (in particular, by presenting the event as a success for the Obama administration), Impact, and Timeliness, the UK press strongly underlines the news values of Negativity (in particular, in The Telegraph), Impact, Timeliness/Superlativeness; finally, the US press strongly enhances the news values of Impact (specifically, focusing on future consequences), Eliteness (in this case, the judges involved in the US Supreme Court ruling were particularly foregrounded), and Negativity (explicitly linked to the opposing sides in the judges’ decision).

Thus, by extending our analysis to the entire week after the ruling and by combining a quantitative and a qualitative analysis of our data, our aim is to highlight if these particular news values are further enhanced by the newspapers under investigation and if the signification spiral portrayed and entailed by playing on these particular (or further) notes of newsworthiness are consistently used by the news media agencies in reporting the event.

References


"Please accept my appreciation": A corpus-pragmatic investigation of thanking behaviour in British and American emails
Rachele De Felice (University College London, UK) and M. Lynne Murphy (University of Sussex, UK)

Previous work on expressions of thanks in English (Eisenstein and Bodman 1993, Aston 1995, Aijmer 1996, Schauer and Adolphs 2006, Wong 2010, Jautz 2013, Mosegaard Hansen 2016) has identified several functions beyond showing gratitude, such as mitigating the negative face threat of a rejection or dismissal (*thanks, but we won't need you*), or part of indirect requests, in the form of prospective thanks (*I’ll thank you to..., thanks in advance for*...). Expressing thanks can be routine phatic work, as noted for example by Eisenstein and Bodman (1993:66)'s description of *thank you* used in American English as "social amenity rather than a genuine expression of gratitude", and Leech (2014:197) who observes in British English that "a large proportion of thankings, as of apologies, are cases where the utterance is highly routinized".

The increasing number of studies on pragmatic variation (e.g. Flöck 2011, Goddard 2012, Haugh and Schneider 2012, Murphy & De Felice in press) indicate significant differences in pragmatic behaviour across “native” English varieties. The small amount of work that has compared thanking in British English (BrE) and American English (AmE) has found differences in how expressions of gratitude are used. BrE, for instance, has been claimed to use *thanks* and *thank you* for additional non-gratitude functions that are not exploited in AmE: to mark points in an exchange (Hymes 1971) and to emphasize disapproval (Algeo 2006). Given that BrE seemingly has more functions for *thank*(s), one might predict that *thank* would be more common in British English. However, AmE speakers use *thank* far more than BrE speakers in corpora of spoken English (Biber et al. 1999) and web-based English (Davies 2013).

The present study offers a systematic examination of thanking expressions in British and American corporate email corpora (the Enron and COBEC corpora: Styler 2011, De Felice & Moreton 2015), echoing previous research on *please* in the same corpora (De Felice & Murphy 2015, Murphy & De Felice in press). In these corpora, we search for a range of relatively direct thanking expressions, including: *thank*, *grateful*, *gratitude*, *appreciat*, *cheers*, *ta*, *thx*. The emails are then coded for rate of thanking expressions, while individual thanking utterances are blind double-coded on several dimensions, including:

- form of expression (*thanks, thank you, I appreciate*, etc.)
- intensification (e.g. *thanks very much*)
- presence / absence of direct address (*thank you, Ruth*)
- recipient (*2nd person you, 3rd person, e.g. thanks to the Alpha team*)
**Preliminary analysis of 8729 emails confirm the “Americans thank more” claim for the business email genre: 49% of American emails include thanks versus 31% of British ones. However, the data also show that BrE displays a broader range of thanking expressions compared to AmE. For example, phrases using *appreciate* constitute 21% of BrE instances, but only 10% of AmE ones.**

It has been proposed (Biber et al. 1999, Murphy 2016) that the disparity between British and American thanking rates may be attributed to a greater positive-face orientation (‘solidarity politeness’) in the US. In line with this, Tottie (2002) claims that Americans are more likely than the British to thank others for their time. Our data shows that, in this context, *thank you/thanks for your time* is a rare occurrence in both corpora; however, AmE writers are much more likely than BrE ones to thank their interlocutor for their help or assistance, both prospectively and retrospectively. BrE displays a wider range of actions and events for which gratitude is expressed, often pointing to material and retrospective actions (e.g. sending and reviewing documents, attending meetings, producing information).

These differences in the object of gratitude and related prospective / retrospective usage could be at least in part accounted for by the use of (and stereotypes about) *please*, the other canonical “magic word” of politeness in these two varieties. Anecdotally, speakers of one variety are alert to differences in use of these words in the other variety, cf.

(1) I often complain that Americans rarely say “Please” but boy do they take “Thank you” seriously (British expatriate in the US; [http://pondparleys.blogspot.co.uk/2011/10/americans-brits-always-offending-each.html](http://pondparleys.blogspot.co.uk/2011/10/americans-brits-always-offending-each.html))

(2) Yet while *thank you* is still important to civilized discourse, I find that *please* has almost the opposite effect in American English. ([http://dialectblog.com/2012/05/13/impolite-please/](http://dialectblog.com/2012/05/13/impolite-please/))

Recent work on corporate emails (Murphy and De Felice in press) has confirmed these impressionistic accounts, reporting a lower use of *please* in AmE together with a greater variety in the phrasing of AmE requests compared to BrE. We suggest that *thanks/thank you* occurs more in AmE because it acts as a polite request marker in lieu of *please*. This may also help account for why the form of thanking is more invariant in AmE.

Further differences emerge from the written nature of the data. For example, while in spoken BrE, Aijmer (1996:59) notes that thanking is “almost mandatory” in closing business telephone conversations, in the BrE corpus *thanks/thank you* is far more likely to open than close an email. The
opposite is true in the AmE data, where, for instance, around 75% of instances of thanks/thank you are found in the closing of the email.

The analysis allows us to address the question of whether stereotypes of a greater positive-face orientation in AmE are warranted across interactional genres. It also highlights the interconnected role of politeness markers seemingly associated with different speech acts, which is more clearly identified by corpus investigation.

References


Developmental Patterns of Metadiscourse in Second Language Writing
Yuichiro Kobayashi (Nihon University, Japan)

The understanding of language learners’ developmental patterns is one of the central issues in second language acquisition (SLA) research. The availability of computer learner corpora allows SLA researchers to focus afresh on descriptive facets of interlanguage analyses and to identify in an increasingly meticulous manner the language characteristics at different developmental stages (Tono, 2013). However, learner-corpus based developmental studies have concentrated upon the lexical aspects of learners’ production (Meunier, 2015), and in turn, have given little attention to second language (L2) discourse development.

In order to address the lacuna of previous scholarship, the present study aimed to profile the developmental patterns of discourse in L2 writings among different first language (L1) groups. Applying the list of metadiscourse markers proposed by Hyland (2005) to learner language, this study investigates variation of metadiscourse across proficiency levels, as well as across L1 backgrounds. This study draws on the International Corpus Network of Asian Learners of English (ICNALE) (Ishikawa, 2013), which is considered to be the largest Asian composition database. The data analysed here is a subset from this database, including L2 writings of six L1 groups (viz., Chinese, Indonesian, Japanese, Korean, Taiwanese, and Thai). In this dataset, all learners were classified into three levels of the Common European Framework of Reference for Languages (CEFR): A2 (Waystage), B1 (Threshold), and B2+ (Vantage or higher). The writing conditions were rigorously controlled for the contrastive analysis of these groups. All compositions in the subset were written in response to a single writing topic, namely “It is important for college students to have a part-time job.”

Starting from the presumption that a “unique matrix of frequencies of various linguistic forms” characterizes every interlanguage (Krzeszowski, 1990), the present study compared the frequencies of metadiscourse markers used in the writings among different learner groups. As correspondence analysis provides a statistical summary of the characteristics of variation, it can be utilized as a first step to consider which metadiscourse features should be investigated in more detail. After identifying analysis points for further investigation, the present study tracked the frequency change patterns of metadiscourse features across proficiency levels in each L1 group.

The results suggest that the six learner groups that were compared have diverse frequency change patterns of metadiscourse features across proficiency levels. To be specific, Japanese learners’ heavy reliance on self-mentions and boosters is remarkably antithetical to Thai learners’ preference of engagement markers and hedges. Moreover, B2 and higher level learners in China and Taiwan exhibited greater numbers of evidentials than learners in other groups. These differences can be attributable to their L1 rhetorical strategy, not to their lexical and grammatical competence. Therefore, we should consider the idiosyncrasies in metadiscourse of each L1 group when assessing L2 learners based on their language
performance. The present study makes a contribution to our understanding of the nature and characteristics of L2 metadiscourse variation.

References


‘When is a metaphor not a metaphor? - An investigation into lexical characteristics of metaphoricity amongst uncertain cases’

Katie J. Patterson (Austral Universidad, Chile) and Michael Pace-Sigge (University of Eastern Finland, Finland)

Introduction

This paper explores the ways in which language users make sense of metaphoricity when manifest in a variety of ways within the language. In order to do this, the research provides an analysis of the lexical characteristics of a set of keywords - grew, cultivate, flame - in a nineteenth century corpus occurring in potentially, but not clearly identified, metaphoric contexts. The analysis focuses on flexible patterns of meaning and the relationship between metaphor and other aspects of figurative language such as polysemy, metonymy and meronymy. The research follows a larger corpus-driven study which found differences in the lexical behaviour of clearly defined metaphoric and non-metaphoric instances of these items when looking at a large set of collocations, colligations, and semantic, pragmatic and textual associations (Patterson, 2017). These behaviours or patterns are consequently avoided by the non-metaphoric instances of that same item, in order to avoid ambiguity. In the case of more ambiguous or unclear cases of metaphor, this paper aims to determine if these patterns are still visible and the extent to which they signal metaphoricity. Evidence of such patterns would imply that lexical, grammatical, textual and pragmatic manifestations in language play an important role in distinguishing between subtleties in word senses and meanings, even in the case of less obvious metaphoricity. As a consequence, awareness of these behaviours or characteristics (or lexical primings) should be at the forefront of any lexical metaphor theory.

Research background and aims

Research into lexical metaphor, within a range of disciplines, has provided well-documented evidence that calls for a rejection to the dichotomist stance. The introduction of corpus methods has, for instance, led to more usage-driven approaches, addressing the sociolinguistic and interpersonal contexts in which metaphors are used (Deignan & Semino, 2010; Deignan, Littlemore & Semino, 2013). This has led to an increase in research acknowledging varying levels of conventionality, strength and salience (Giora, 2003, Svanlund, 2007; Steen et al., 2010; Deignan, Littlemore & Semino, 2013). However, whilst deriving metaphoric data from corpora is by now well established within the field (for example Partington, 1998; Deignan, 2005, Koller, 2006), its premise of focusing on
repetitive patterns of use means that some cases of metaphoricity are often ignored. Corpus studies concentrate solely on metaphors at either end of a frequency cline. These are the often-repeated, conventional instances (those that conform to the pattern), or the novel and highly creative ones (those that exploit the pattern). In such studies, little focus is placed on the problematic instances, or the types of language of which readers remain unsure or hesitant to clearly identify or label.

The intention of this paper is to focus on levels of metaphoricity in corpus data, not from the notions of originality or conventionality, but instead from the individual language user and their interaction with individual, problematic instances of metaphor within a given context. The aim is to highlight the importance of lexical, grammatical, textual and pragmatic manifestations of meaning and their role in distinguishing between subtleties in word senses and uses. The hypothesis is that these subtle relationships, such as the collocates a word has in a given use, its grammatical constructions and its pragmatic effects, all help us to determine which sense of a word is being meant in a given context. Furthermore, it argues that this can be extended to ambiguous forms of language, where a distinction between word senses is not clearly defined.

Previous research has documented that a metaphor’s ability to violate or bend the limits of linguistic conventions (semantically, lexically, grammatically), provides it with its freedom in language (Deignan, 2005; Philip, 2011; Patterson, 2016; Patterson, 2017). Moreover, the flexibility of figurative language is what gives it its pervasiveness in everyday communication: we are able to manipulate it so well to our own needs. This study aims to investigate the extent of the ‘fuzziness’ around the unsure or problematic cases in order to determine if creative language is as unrestrictive as it claims it to be. The research aims can be formed as follows:

1. Do any of these problematic instances of metaphor share any defining lexical features?
2. What do the results tell us about where metaphorcity exists in the language?
3. What does this mean for our understanding of metaphor?

**Methodology**

The corpus consists of texts written by English authors between 1800 and 1899. In total, there are 416 texts with a running token size of 45,480,658. WordSmith5 (Scott, 2009) was used to extract the three chosen words for this study from the corpus. An initial Keyword search identified words of unusually high frequency in the nineteenth century corpus in comparison with a more general and contemporary comparator corpus (the BNC_W). The lexical item approach is the same as the one taken by Lindquist and Levi (2008), and the opposite of the standard approach of many studies on metaphor, “which tend to start from a particular semantic field” (Lindquist and Levin, 2008: 145). This allows for an exploration of all possible uses of an item in a variety of behaviours
and does not single out a particular type of metaphor, based on a single feature or characteristic. Moreover, it accounts for phraseological manifestations of meaning and possible idiomatic uses. Where a key item is singled out methodologically, the analysis will be exhaustive of all the item’s occurrences and more importantly, will concern co-textual as well as contextual and text-linguistic features. Each item will be studied primarily within the framework of its concordance line.

The chosen keywords were passed on to a group of raters who identified their use (within their given concordance line) as either metaphoric, non-metaphoric, or unsure. The aim was to acquire two clear-cut groups of metaphors and non-metaphors whose items have all been unanimously decided on, to reduce subjectivity. If raters came to different decisions about instances, these instances were automatically placed in the middle group. This middle group is the focus of this study. The items and their corresponding concordance lines in this unsure, middle group were turned into a mini-corpus for this investigation.

Examples of such unsure metaphors include references that are potentially ambiguous:

(1) There was a road there once, perhaps, when Cundinamarca was a civilized and cultivated kingdom.

In this example, it is unclear if the instance refers to the physical landscape within the kingdom, or the abstract whole, in which case the meaning would be metaphoric. Other instances are potentially metonymic:

(2) new centres of population grew up at a distance from the original village.

Here the use of the lexical item grew up rather than grew, implies a maturing associated with the individuals rather than the physical population size. Other instances are potentially polysemous:

(3) The old gentleman grew infirm.

In this example, as with many others, grew can be interchanged with became. There are also highly conventional metaphors in this middle group:

(4) the shrieks of death that wildly broke through devouring flame and smothering smoke.

In this example the decision on metaphoricity rests on the extent to which a flame can devour something. This raises questions relating to animacy and animate adjectives in particular, which are commonly associated with a flame, such as consuming and raging.
Analysis and results

Unclear or potentially metaphoric instances of the keywords *grew*, *cultivate* and *flame* together with their concordance lines, form three separate sub-corpora. These are then analysed, quantitatively and qualitatively, for evidence of lexical characteristics. The results show the item to behave along a cline of metaphoric behaviour, integrating word senses such as polysemy and metonymy. In this sense, the research works outwards from the language user, in the true sense of corpus-based studies, and draws conclusions both lexically and cognitively, from the individual instances of language. The findings are also compared and contrasted to contemporary and spoken corpora to determine if these patterns of use of dependent on the the time period and text type of the corpus.

The study shows that lexical characteristics were found to be present amongst sets of instances: a finding which has implications on our notion of a metaphoric cline. Whilst metaphoric language remains creative to some extent, identification at the level of the lexis or beyond suggests all types of metaphor to be more formulaic than previously thought. The finding also suggests that metaphoricity can be identified through the presence or absence of lexicogrammatical markers. This impacts on both theories of metaphor identification and metaphor understanding.

References


Drought and the scarcity of water pose substantial challenges to the natural environment within Britain and throughout the world. In order to improve our ability to manage instances of future drought, we would benefit from a greater understanding of the causes and impacts of droughts of the past. Droughts are not necessarily caused by lack of rainfall alone and can be worsened by a number of socio-economic factors such as the rate of water consumption. By increasing our knowledge of attitudes to water usage and scarcity, we can foster more informed public dialogues about issues concerning lack of water.

The UK now has an Environment Agency which monitors instances of drought and acts to reduce their impact but, prior to its establishment, there was no national system which recorded occurrences of drought. Droughts were, however, often mentioned in both national and regional newspapers throughout the nineteenth and twentieth centuries. In this paper, we examine two hundred years of newspaper data, from 1800 to 2014. We aim to quantify and characterise occurrences of drought and water scarcity throughout the United Kingdom and the rest of the world, and explore British public discourses surrounding drought.

One of the initial challenges we faced in gathering the data was identifying an appropriate set of search terms that could be used to find references to droughts. By using query terms which excluded surnames and other irrelevant words, we found that references to drought overwhelmingly related to lack of rain within the nineteenth-century news texts. However, this was not the case when we considered material from the twentieth and twenty-first centuries due to a common tendency to use drought as a metaphor, particularly in journalism relating to sporting achievements (e.g. goal, trophy or scoring droughts) and finance (credit, mortgage droughts). Indeed, in these modern texts we found that the vast majority of references to drought were not presented literally. This meant that we were required to manually clean texts where necessary in order to isolate and exclude metaphorical references to drought.

Another challenge was the fact that newspapers do not discuss drought within the context of Britain only. As we were primarily interested in examining the discourses around drought events in Britain specifically, identifying where the drought happened was a key issue in the analysis. Here, we have utilised an innovative and developing methodological approach which combines corpus linguistics with GIS (Geographic Information Systems) methods, thus adding a useful geographical element to our analyses. In the first instance, this will allow us to visualise collections of data with ease and prove to be of particular benefit when dealing with very large amounts of raw data. By uncovering spatial patterns in types of discourse we are able to ‘think geographically’ and explore geographically bounded discourses through the examination of the link between textual patterns and geographic references. We have applied a process called concordance geoparsing (Gregory and Hardie 2011, Rupp et al. 2016) to our dataset. This
technique is achieved by means of a number of stages. Firstly, we used corpus linguistics software to extract each occurrence of a search-term and a span of fifty words to the left and right of it. Secondly, these concordances were geo-parsed in order to identify instances of place-names, for which co-ordinates were then found. The results were manually analysed in order to reduce errors. In essence, concordance geo-parsing allows us to find all instances of place-names which co-occur with a search term of our choosing. The co-ordinates we generated were then applied to GIS software for mapping.

Concordance geo-parsing allowed us to divide our data into three broad categories: i) droughts which are reported as occurring in Britain; ii) droughts which are reported as occurring outside of Britain; and iii) references to drought which are mentioned in a more general context, for example in relation to climate change or a breakthrough in research. Each set was analysed by a combination of quantitative and qualitative methods, using standard corpus methods – collocational analysis of selected key words – and closer reading of the concordance lines.

We found that drought and water shortage are mentioned in our dataset in rather vague geographical terms, such as occurring across England or the UK, even though droughts seldom affect an entire country – some affect large parts of a country but others are limited to a relatively small area. Within the nineteenth century, droughts were often mentioned as impeding localised agricultural output. However, writers did associate lack of water with the spread of infectious disease among the population of Britain, particularly in children. In more recent years, drought has taken on a wider frame of reference. Droughts are increasingly mentioned alongside other extreme weather events and in relation to climate change. Writers also show awareness of global issues in referencing drought, particularly those affecting the continent of Africa. Alongside exploring public discourses surrounding drought, we will show how attitudes to water shortage have changed over a period of two hundred years.

References


Using Multidimensional Analysis to Investigate the Discussion Sections of Research Articles in Chemical Engineering

Bixi Jin (The University of Hong Kong, Hong Kong)

Introduction

Driven by the ‘publish-or-perish’ academic culture, there is increasing pressure on scientists and postgraduate science students to publish in prestigious high-impact journals, either for the dissemination of research recognized by international academic community or for professional advancement and promotion. In the process of writing up research articles, previous studies have shown that writing the discussion of results section can be a particular challenging undertaking for EAL novice scientists (e.g. Bitchener and Basturkmen, 2006). This ongoing study investigates the variation in underlying communicative functions in Discussion sections of high- and low- impact research articles in engineering. It is attested in the literature that no comparisons have been made in this aspect. The comparison in this sense is important as it can directly make less experienced scientists familiarize with and aware of the prevalent discourse patterns of discussing research findings in these ‘successful’ and ‘less successful’ models of discussion of findings, so as to navigate their participation in international academic discourse community.

Corpus compilation and analysis

A total of 213 RA Discussion sections in chemical engineering were searched and compiled for the analysis based on citations and journal five-year impact factors. One sub-corpus consists of 113 discussion sections coming from the highly cited research articles of high-impact journals published over 2005 to 2015. The selection was to ensure that the texts in this corpus have good quality both in terms of language and content, and thus can be regarded as "situationally effective and the results of expert performance” (Bazermen, 1994, cited in Yang, Zheng and Ge, 2015). The other corpus was made up of 103 discussion sections of research articles with few citations in the less recognized peripheral journals. It should be noted that the selected texts are limited to a single discipline (i.e., chemical engineering in this study), as there might be potential differences arising from the different disciplines.

Multi-dimensional (MD) analysis developed by Biber (1988) was used as the major approach for comparison by looking at the interaction among a range of linguistic features, as single linguistic features cannot reliably account for the description of variation in language use (Biber and Conrad, 2001). Using MD to investigate variations of part-genre of research article has been rare, with the exception of Xiao and Cao’s (2013) study of abstract written by NS and NNS writers. Multi-dimensional Analysis Tagger (Nini, 2014) was adopted for automatic tagging. The tagged features were then normalized to per 100 words considering the average length of the individual texts. Principle Component Analysis (PCA) was then conducted on the normalized features using SPSS 22.0. The six-factor solution was determined according to the observation of the scree plot. The first six factors account for 39.284% of the total variance.

In order to look into the significant differences of RA discussion sections in different types of journals, the mean scores of each dimension were computed using z-scores for the two corpora and then subjected to an Independent Sample T-test for analysis.
Preliminary findings

The PCA analysis reveals the six underlying communicative functions of discussion sections in the Corpus of High-impact articles (CHA) and Corpus of Low-impact Research Articles (CLA). The six dimensions with proposed descriptive labels are:

Table 1 Composition of the positive and negative features of Dimension 1-6

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Linguistic features and factor loadings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dim 1: Involvement and interactivity</td>
<td>Split auxiliaries (0.746), total adverbs (0.725), amplifiers (0.485), modals (0.485), conjuncts (0.428), perfect aspect (0.408)</td>
</tr>
<tr>
<td>Dim 2: Narration vs. non-narration</td>
<td>Present tense (0.848), pronoun it (0.449), that-deletion (0.342), place adverbials (0.312), past tense (-0.779)</td>
</tr>
<tr>
<td>Dim 3: Further explanation and elaboration on evaluative expressions</td>
<td>Predicative adjectives (0.856), be-verb (0.847), subordinators (0.503), demonstratives (0.402)</td>
</tr>
<tr>
<td>Dim 4: Informational density</td>
<td>Average word length (0.686), attributive adjectives (0.683), do (0.486), common nouns (0.454)</td>
</tr>
<tr>
<td>Dim 5: Stating results/claims</td>
<td>that-complements (0.698), private verbs (0.615), public verbs (0.458), first-person pronouns (0.4)</td>
</tr>
<tr>
<td>Dim 6: Expression of denial relationships in experimental findings</td>
<td>Negation (0.642), existential there (0.496), to-infinitives (0.477), prepositional phrases (-0.408)</td>
</tr>
</tbody>
</table>

Dimension 1: Involvement and interactivity
Dimension 2: Non-narration vs. narration
Dimension 3: Further explanation and elaboration on evaluative statement
Dimension 4: Information density
Dimension 5: Stating results/claims
Dimension 6: Expression of denial relations towards statement or experimental findings

The follow-up independent sample T-test compared and contrasted the differences of discussion sections in two types of journals along these six dimensions. Dimensions 1, 3, and 5 are found to be significantly different in two corpora.
The T-tests show that RA discussions in CHA appear to discuss their results by employing a constellation of linguistic features associated with interactivity and involvement, while this is not the norm in the CLA. That is, authors in CHA are capable of using more metadiscourse devices like hedges, modal verbs and boosters to manage a good relationship with readers and make their claims acceptable by their disciplinary communities (see Hyland, 1996; Hyland, 2005; Yang, Zheng, and Ge, 2015), whereas authors in CLA use metadiscourse-related features in a less frequent manner. The contrasting of using hedging devices in CLA is consistent with a series of findings (see ElMalik and Nesi, 2008; Hu and Cao, 2011; Yang, 2013). The possible explanation for this could be related to influence of two different scientific communities (international vs. local), as the majority of the discussion sections of low impact articles included in the subcorpus are national-based English journals (e.g., the Chinese journal of chemical engineering, the Korean journal of chemical engineering, etc), which attract a fair number of researchers in these countries and regions to publish in them, though the journals aspire to attract more international audiences. The research writers from these local scientific communities, as suggested by Loi, Lim, and Wharton (2016), are inclined to display a less reader-friendly style of writing. That is, they do not invite readers to be actively engaged in research, but require them to participate in and interpret the intention and argument on their own.

**CHA and CLA on Dimension 3**

In Dimension 3, the result suggests that the preferred communicative styles of discussing results in CHA often constitute a evaluative sentence followed by elaboration and explanation on this statement by making using of adverbial subordinators such as if, because, though, whereas, and since to indicate the relationship of time, reason, condition, and comparison. On the other hand, discussions in CLA make fewer comments on research findings, and have less concern for details of the use of these corresponding
adverbials that makes their evaluation of findings specific. Similarly, the reason could be partly related to the fact that writers from local scientific communities are less likely to adopt a stance in their work by using value-laden words, making it “more prominent and appealing” (Loi et al., 2016, p.12) due to the less competitive local publishing context.

**CHA and CLA on Dimension 5**

A final interesting point of difference is in Dimension 5. The discussions in CHA display obvious authorial presence to make explicit claims by making use of the complement clauses controlled by verbs and also the first-person pronouns. Discussions in CLA have a tone of objectivity that avoid direct referring to themselves, but give explicit reference to the research results (e.g., *these results indicate...., our results implied...*). The revealed difference partially aligns with Hyland’s (2002) finding that novice writers appear to conceal their role due to a lack of confidence, they “consciously avoided the most authoritative functions and sought to deny ownership and responsibility for their views” (p. 1107). In contrast, expert writers are more inclined to make knowledge claims to establish “a personal authority based on confidence and command of their arguments” (p.1104). Although no absolute claims can be made that discussions of high-impact journals are all written by expert writers, these successful models can be considered as ‘expert performances’ (Bazerman, 1994), and can thus partially allude to Hyland’s (2002) findings that suggest the use of first person pronouns to “speak as an authority” is important for successful academic writing (p.1094). Also, the findings may also point to the possibility that maintaining objectivity may still be a rooted mindset for local scientific communities.

**CHA and CLA on Dimension 2, 4, 6**

In contrast, no significant statistical significance can be found in Dimensions 2, 4 or 6 and thus are not discussed with further details.

**Conclusions**

The corpus-based comparative study addresses two research questions, i.e., to describe the underlying functional dimensions of 213 RA discussions, and to explore the linguistic variation of discussion sections of high- and low-impact RAs. Six dimensions using Principle Component Analysis (PCA) were generated and three of them were used as the focal aspects for comparison. The significant variation found in Dimension 1, 3, and 5, suggesting that successful models of RA discussions incorporated proportionally more metadiscursive features, first-person pronouns and evaluative statements with further explanation. The differences may be related to the ‘expert’ and ‘non-expert’ performances exhibited by two groups of researchers.

The contrastive findings have pedagogical implications for novice scientists and EAP practitioners. Less experienced scientists can adapt their writing of this section to RA discussions of high-impact articles published in high-profile journals, thus enabling their work to achieve greater international visibility. EAP writing instructors can make use of the contrastive findings to develop corpus-informed teaching materials, and through a series of consciousness-raising tasks, help novice writers notice the linguistic characteristics underlying successful RA discussions and skillfully manipulate linguistic patterns to produce convincing and compelling claims in their manuscripts.
References


Data-Driven Learning (DDL) has been increasingly used in the technical writing classroom as an effective method for introducing general- and discipline-specific language features to learners (Boulton 2012). One of the many strengths of DDL over traditional prescriptive approaches to writing instruction is that it allows learners to utilize or construct discipline-specific corpora that meet their unique target language needs. As a result, even within a heterogeneous class, each learner can discover the characteristic features of their own discipline, and through discussion and collaboration with others working with different corpora, learn about discipline specificity (Anthony, 2012), interdisciplinary features of writing (Bhatia, 2010), and general patterns of language use. However, DDL also suffers from a number of weaknesses. In particular, there is a general lack of discipline-specific corpora (Ädel, 2010). This means that the instructor or learner must construct an individualized corpus, but this requires locating, downloading, cleaning, and tagging relevant language data that is often beyond their technical skill level. In addition, searching within an individualized corpus for discourse-level language features is extremely limited due to the lack of any annotation.

In this research, the lack of availability of suitable discipline-specific corpora is addressed through the development of AntCorGen, a freeware corpus generation software tool that can automatically search, collect, clean, tag, and annotate very large discipline-specific corpora. The AntCorGen tool runs on all major operating systems. It is standalone and portable, and requires no installation or security permissions. Also, the tool is designed to be simple and easy to use without the need for instruction guides or tutorials. These features make it ideally suited for use by instructors in the creation of DDL class materials and by learners as part of in-class DDL learning tasks and activities.

AntCorGen utilizes the PLOS API [1], which provides access to the PLOS ONE multidisciplinary Open Access journal [2]. Through a user-friendly graphical interface to the API, users can search for relevant articles in the journal database using broad or very narrowly-defined parameters including subject category, keywords, date of publication, popularity, and type of article. The tool then automatically downloads relevant articles that match the search criteria and stores the data on the user’s local file system in plain text and PDF formats. This allows the data to be immediately viewed in its published form or analyzed with traditional desktop corpus tools, such as AntConc (Anthony 2017). Using the metadata included with the articles returned by the API, AntCorGen can automatically divide the texts into sections and annotate this data accordingly, saving the data as separate files and placing them in folders that are named according to their headings. This function makes it simple for a user to search for language patterns within, for example, the title, abstract, introduction, materials/methods, results, or discussion sections of a research article.
To evaluate the utility of AntCorGen, 236 research articles in the area of "Human Mobility" and 126 articles in the area of "Temperate Forests" (both subject categories within PLOS ONE) were collected using the tool. Word lists and p-frame lists were then generated for each rhetorical section category in both subject areas using AntConc (Anthony, 2016) for the word list generation and AntGram (Anthony, 2017) for the p-frame generation. Using these lists, learners are able to investigate the linguistic features of each section of a research article through corpus analysis. For example, in the "Human Mobility" data, learners can find that the frame <that the # of> is relatively frequent in both the introduction and discussion sections, and the frame <we # that the> is relatively frequent in the "Temperate Forests" discussion sections. Such observations raise questions about why certain words and p-frames are more frequent in one section over another, which can lead learners to a deeper analysis of the relevant concordance data. In the case of <we # that the>, for example, learners can find that half of the instances are reports on what the researchers found, demonstrated, or showed, while the other half utilize present simple verbs and congregate around acts of proposing, hypothesizing and suggesting (e.g., <we suggest that the>). Having seen that we is used in the discussion section in this way, the learners can then investigate whether or not we is used in the same way in other sections of a research article.

In conclusion, the AntGram corpus generation tool presented here makes it possible for learners to create individualized discipline-specific corpora for use in DDL, which have the added facility of subdivision of the research articles into principal rhetorical sections.


References


Basic corpus annotation made easy: The Language Analysis Portal (LAP)
Jarle Ebeling, Emanuele Lapponi and Milen Kouylekov
(University of Oslo, Norway)

This poster presentation describes and showcases the CLARINO Language Analysis Portal (LAP, http://www.mn.uio.no/ifi/english/research/projects/clarino/), developed and maintained at the University of Oslo. LAP was launched in September 2016 after having been under development since 2013 (see Lapponi et al. 2013). Researchers interested in using the tools within LAP, can log in using their CLARIN or eduGAIN user account.

LAP is a user-friendly web interface to common annotation tasks such as tokenization, POS tagging and syntactic parsing of naturally occurring text. Basic annotation of texts is nearly always necessary as the first step in creating a corpus resource, whether it is intended to be a searchable resource or used as input for further analysis, e.g. semantic annotation. However, the tools available to achieve even the most basic annotation often demand non-trivial programming skills or installing and running different pieces of software one after the other from the command line. LAP helps the so-called "non-command-line-proficient researcher" to overcome these obstacles by letting the user drag, drop and link tasks in the web interface to create a workflow. The workflow can be stored and re-used on other data sets and shared with other users.

LAP is integrated with the Galaxy workflow system (https://galaxyproject.org/) to allow users to specify and run a series of tasks, known as a workflow, e.g. segmentation, tagging and parsing, and then download the output from all, or only the final of, these tasks. Originally developed for biology, Galaxy has been adapted to a range of other research areas, e.g. linguistics and social sciences. At the University of Oslo there is a team of developers devoted to maintaining and developing similar workflow systems for the life sciences, geology and metrological data in addition to linguistics and the social sciences.

LAP runs on a high-performance computer, handles large data sets, and recognises several languages, e.g. English, Norwegian and Sami. The members of the LAP team do not develop NLP processing tools themselves, but rather implement and make state-of-the-art (NLP) tools available to end users in an easy-to-use interface (Lapponi et al. 2015). The poster presentation will illustrate some of these tools in action.

The LAP development team welcomes suggestions for further open source tools to be integrated in the portal.

References

Populism in Italy has been inscribed for a long time in the right-wing political tradition. However, in recent years, a new form of populism has emerged, one that somehow overcomes the long-established binary opposition between right and left. Indeed, next to a more conventional form of Italian populism, such as the one envisaged and promoted by the Lega Nord party, the Italian “Movimento 5 Stelle” (5 Star Movement - M5S henceforth) defies and challenges the status quo of the Italian political scene, which traditionally revolved around a bipolar, yet fragmented, rationale.

In 2009, when many European countries saw the rising of far right-wing populist parties, the Italian comedian Beppe Grillo and the web strategist Gianroberto Casaleggio started the M5S, which rapidly gained consensus among the electorate. From its outset, M5S attracted voters from very different social categories, creating in this way a wide web-based core of followers, using the blog as the main hub for their claims.

In 2013 general elections, M5S managed to be represented in Parliament as opposing party. Finally, in 2016 local elections the Movement won two important urban centres (Rome and Turin), securing the opportunity of full institutionalisation and entering the establishment, which M5S has always tried to demystify and challenge.

The most peculiar characteristic of M5S populism is precisely its ability of bringing together people from different social and professional categories, whose positions would have been at opposite poles in the tradition of Italian politics.

Since the blog is M5S’ main method of addressing the electorate, we built a corpus of texts from www.beppegrillo.it and analysed the ways this new kind of populism addresses its own electorate. The M5S corpus gathers blog posts by activists and political representatives of the Movement from April to July 2016, comprising 521 texts and approximately 240,000 words. Our goal with reference to the M5S corpus is to analyse the Movement’s approach and definition of both its own identity as a political entity and its relationship with (prospective) voters. Following the corpus-based approach within the CADS framework (Partington et al., 2013), we looked at collocates and concordance lines of words¹ such as noi (we), cittadin* (citizen*), popolo/gente/person* (people), in order to see how M5S relates to these entities and how, consequently, negotiates its identity and political influence.

Our first findings showed a certain inconsistency on the part of M5S in building and negotiating its political identity, which is defined mainly through three different and recurring communicative strategies.

Firstly, we observed M5S’ tendency to define itself, i.e. the ingroup (van Dijk, 1998), as a political movement by means of opposition to and negation of “traditional” political and social forces, i.e. the outgroup (van Dijk, 1998). Indeed, the negative adverb non (not) occurs among the top collocates of both the pronoun noi (we) and

¹ Sorted by T-score, word span -/+5 words.
the verb *siamo* (to be, indicative, present, first person plural). This seems to hint at its need of denying an already existing system (Italian political establishment) to strengthen and affirm its political self.

Secondly, looking at the concordance lines and collocates of the word *cittadin* (citizen*), we identified two recurring patterns. The first one is the polarized relationship between *cittadin* and the establishment (banks, politicians, government etc.), where the latter is a negative force upon society, and citizens are framed as good and honest. The second pattern represents citizens as weary of this status quo, and calls for them to take an active stance. In this context M5S’ representatives, while carefully detaching themselves from traditional politics, allineate themselves with citizens through the frequent pattern *noi cittadini siamo* (we [as] citizens are) and addressing their audience with compassionate and paternalistic terms. Therefore, to the eyes of its followers, M5S represents the only political and social force that lets citizens regain an active role in society and empower themselves against the establishment, which is depriving them of their fundamental rights (proper income, house, retirement, etc.).

Thirdly, we looked at three words (*gente, popolo, persone*) that can be conveyed by the English word “people”, thus expressing a collective meaning. In particular, *persone* is used in two main ways. On the one hand, it has a factual meaning when collocating with numbers. On the other hand, *persone* has an evaluative function, which emerges also in the word *gente*, with which it shares some collocates. These collocational patterns shared by *gente* and *persone* point at two different and opposing attitudes towards people: again, a paternalistic attitude conceivably referring to those who are part of M5S’ ingroup (party members, followers and voters), this attitude is expressed through the adjectives *brava/e* (good), *povera/e* (poor); and an attitude of detachment (e.g. *certa/e* - certain, some) employed ironically or argumentatively to refer to people outside the movement who support or are part of the establishment, that is the outgroup. Finally, *popolo* is mostly followed by adjectives describing a geographic/national provenience (*italiano* - Italian, *ucraino* - Ukrainian, *turco* - Turkish) or a status (*sovranità del* - sovereignty of, *governo legittimato dal* - Government deriving legitimacy from). The *popolo* ideally identifies with M5S followers and is depicted as victim of some (mainly establishment-related) perpetrators.

The trends we observed in the M5S corpus seemed to highlight some basic traits that M5S shares with more conventional populist parties, i.e. criticism towards traditional politics, seen as threatening to people’s rights and as reducing citizen’s sovereignty (Biorcio, 2014). Moreover, following Spiessen and Van Poucke (2016: 323), and van Dijk (1998), we agree that social groups represent themselves through ideology, using “self-serving” schemata that establish a conflict between in- and outgroups, between "us" and "them", in fact:

The very general polarization schema defined by the opposition between Us and Them suggests that groups and group conflicts are involved, and that groups build an ideological image of themselves and others, in such a way that (generally) We are represented positively, and They come out negatively. Positive self-presentation and negative other-presentation seems to be a fundamental property of ideologies. Associated with such polarized representations about Us and Them, are representations of social arrangements, that is, the kinds of things we find better
(equality, a clean environment) or those which we believe others stand for (inequality, a polluted environment, a free market). (van Dijk, 1998: 69)

These "polarized representations" seem to be confirmed by our observations about the M5S corpus. In particular, we believe that the opposition between ingroup (good) and outgroup (bad) is well reflected by Grillo’s populist discourse. From our observations, we were able to suggest that M5S’ identity still relies heavily on establishment’s stigmatization rather than promoting and building their own political identity and agenda.

References


Exploring Grammatical Colloquialisation in Non-Native English
Xinyue Yao (Renmin University of China, China) and Peter Collins (University of New South Wales, Australia)

Colloquialisation, a process by which "writing becomes more like speech", has been identified as a powerful discourse-pragmatic mechanism driving grammatical change in English. To this day, the study of colloquialisation has been largely restricted to native English varieties, in particular British and American English (BrE and AmE). Until recently, possible manifestations of colloquialisation in non-native English varieties have not been investigated in detail. This gap has been partially addressed in a 2014 issue of the Journal of English Linguistics compiled by Nöel, Van der Auwera and Van Rooy. Focusing on expressions of modality in English, the contributions to this issue examine convergences and divergences between Philippine English (PhilE) and its historical input variety, AmE (Collins et al. 2014), and between Black South African English and its native counterpart in the same contact setting (Van Rooy & Wasserman 2014). Further extending our understanding of the evolution of non-native English is the volume edited by Collins (2015), which presents diachronic analyses of several non-native varieties. However, as far as colloquialisation is concerned, the results of existing studies are very mixed and far from conclusive as they focus on the changing patterns of use of only a handful of grammatical features.

In this paper we further discuss the role of colloquialisation in non-native English by reporting the findings of a comprehensive, comparative corpus-based study of PhilE and AmE in the second half of the 20th century. We measure colloquialisation by employing a data-driven approach and relying on the term “colloquiality”, which we use to refer to a combination of the degree of preference for linguistic features more typical of speech (“colloquial features”), and the degree of dispreference for linguistic features more typical of writing (“anti-colloquial features”). Our data consist of corpus texts sampled from the following corpora: (1) the Philippine and American components of the International Corpus of English (ICE-Phil and ICE-US); (2) the Santa Barbara Corpus of spoken American English (SBC); (3) the Brown Corpus, and (4) its Philippine counterpart, Phil-Brown. Together the sampled texts represent two time periods, the 1960s and the 1990s, and a wide range of spoken and written registers. Based on such data, we derived an operationised measure of colloquiality, itself a summary of 18 colloquial features and 8 anti-colloquial features, by conducting ANOVAs for 87 (lexico-)grammatical features that have been shown by existing studies to be implicated in the general speech-writing divide.

To compare degrees of colloquialisation in PhilE and AmE, we applied our measure of colloquiality to a diachronic, parallel corpus of PhilE and AmE. As a
subsection of our entire dataset, this corpus consists of three written registers with distinct situational characteristics, press editorials, learned writing and fiction. An analysis of colloquiality reveals several noteworthy patterns. Regarding register variation, we find that the three registers are not drastically different in their use of colloquial features. Rather, the most remarkable difference on a global level lies in the frequencies of anti-colloquial features, which indicate overall lexical diversity and informational richness.

Regarding diachronic variation, evidence for colloquialisation is register-differentiated. There are considerable increases in the colloquiality scores of PhilE press editorials and AmE fiction over the 30-year period, whereas learned writing does not show remarkable changes irrespective of variety. We argue that the distinction drawn by Hundt & Mair (1999) between “agile” and “uptight” registers is particularly useful for interpreting the diachronic findings. Popular registers which are driven by the need to cater to a large readership tend to be open towards stylistic innovations, whereas specialised registers are less receptive to change with their small and stable audiences. Differences in the nature of the intended audience account not only for the different findings for the two popular registers and learned writing, but also for the contrast between the rapid colloquialisation of PhilE press editorials and the stability of PhilE fiction of the same time period. Importantly, we find that the evolution of PhilE registers cannot be explained by a simple process involving emulation of AmE. We argue that this is inevitable given the unique sociohistorical circumstances in which PhilE has evolved. PhilE’s colonial history imparts to it an elitist character, placing it in a hierarchical relationship with the local languages. The patterns uncovered in this study lend support to the general observation made in previous research that PhilE is less colloquial than AmE. However, there is no convincing evidence for PhilE proclaimed “monostylisticism” (Gonzalez 2004), a tendency to transform features characteristic of formal written English to less formal registers, since stylistic differentiation in PhilE is fairly marked is on the whole.

References

In this article, we report results coming from two sorts of research works. The first involved examining the distribution of items on the Academic Collocation List (ACL, Ackermann & Chen, 2013) in a computer science corpus which contains more than 1,300 articles from high-quality journals. Second, for certain ACL entries which were found to appear rarely in the domain-specific corpus, we investigated whether corpus-derived collocation clusters were able to provide semantically similar alternatives (e.g. academic community as an alternative for the ACL item academic circle). The two sorts of works taken together suggest an innovative approach to collect academic collocations for a particular domain; English for specific purposes (ESP) instructors can first identify which ACL items are frequent (or non-frequent) in a selected domain, and, for non-frequent ones, utilize our collocation-cluster techniques to acquire domain-specific academic usages.

**Distribution of ACL entries in computer science texts**

Ackermann & Chen’s (2013) Academic Collocation List arguably has been the most well-established collocation list compiled for academic purposes to date. Compared with similar works, the ACL is of higher pedagogical value because it adopted a rigid selection process which involved expert judgment. Ackermann & Chen utilized a four-stage approach to extract collocations from the Pearson International Corpus of Academic English (PICAE). The written curricular part of the PICAE, specifically, consisted of texts from four fields of study: applied sciences and professions, humanities, social sciences, and natural/formal sciences, with each covering seven disciplines. To gather candidate collocations, the authors first performed computational analyses using frequency, mutual information (MI), and t-score as main measures. Next, the collected lexical pairs were “refined” with the ones holding lower MI and t-scores or consisting of non-target part-of-speech (POS) combinations (e.g. determiner-noun) being removed. Human intervention was introduced at the third stage during which only the collocations judged as of higher pedagogical use were retained. The last step, systematization, was adopted to make the ACL more systematic and useful. Several function words were added to collocations (e.g. 'be' generally accepted) if deemed necessary. The results of such four-stage process were a 2,468 collocation list. Being examined in the source corpus as well as a general-purpose comparison corpus, the ACL exhibited a 14-times higher frequency in the former, suggesting the high academic nature of the ACL.

Well-established as it is, however, to our knowledge whether the ACL is indeed useful across different professional domains has rarely been investigated or reported. The current study, serving as one of the first to empirically examine the distribution and usefulness of the ACL, focused on computer science, a discipline covered by the written curricular part of the PICAE. Our purposes, as specified earlier, were to explore whether the ACL entries frequently appeared in the selected discipline and which entries were particularly frequent/non-frequent. Totally, our computer science corpus (CSC) comprises over 14 million running tokens coming from texts of high-quality journals. To ensure that the CSC texts are representative of current written discourse in computer science, we used only the journal papers published from 2014 to 2016. The CSC contains articles from...
twelve major computer science sub-domains, including, for example, artificial intelligence and human-computer interactions. For each sub-domain, we consulted four high impact factor journals and extracted about 30 articles from each of them. The collected articles were further “refined” with some unwanted sections (e.g. authors’ affiliations and references) being excluded. We applied Stanford Log-linear Part-Of-Speech Tagger (Toutanova, et al., 2003) on all CSC sentences which then enabled us to check whether the different POS combinations on the ACL showed up in them.

Here we focus on and report non-frequent ACL items in the CSC, as shown in Table 1. We adopted the normed frequency (0.2 times per million) used by the ACL; consequently, an ACL item was considered non-frequent if it appeared only once or never appeared in the CSC. Overall, although we did find that (1) several academic usages (e.g. experimental result, 1011 times) were highly frequent and (2) collocations related to information technology (e.g. natural language, 246 times) or statistics (e.g. statistically significant, 319 times) appeared very often, 25.4% of ACL items were not frequently used in computer science. In terms of different POS combinations, most verb-based pairs (e.g. make available, 201 times, and address issue, 286 times) were high- or medium-frequency ones whereas many adjective- or noun-based collocations appeared rarely in the CSC. One of the reasons for the latter finding is straightforward: many adjective- or noun-based ACL collocations seemed closely related to social sciences. Although Ackermann & Chen (2013) attempted to adopt the ones appearing across disciplines, a certain number of their entries still appeared to be highly humanities- or social sciences-relevant (e.g. capitalist economy and culturally specific). Our results generally support Hyland & Tse’s (2007) claim that “coverall” lexical lists do not reflect the real needs of ESP students and what they need should be a lexical repertoire particularly collected for them.

We in our CSC data also observed some domain-specific word meanings (e.g. value referring to quantity as in high value) and word choice (e.g. using stage rather than phase as in initial stage). These phenomena again support Hyland & Tse’s (2007) suggestion of studying disciplinary rather than general lexical conventions.

### Automatic collection of domain-specific academic collocations

In the second part of this study, we investigated whether it was likely to automatically generate domain-specific academic collocations based on corpus-derived collocation clusters. Our collocation-cluster techniques, developed based on Cowie & Howarth’s (1996) notions of “overlapping collocations”, utilize a hybrid approach taking into consideration both word co-occurrences and semantic information to establish networks of collocations. As two words (e.g. verb-noun pairs such as achieve purpose) are selected, our collocation-cluster system begins to automatically search a corpus in order to identify collocates for both words (e.g. goal/quality/objective for achieve and

<table>
<thead>
<tr>
<th>Numbers of Non-frequent ACL Items in CSC</th>
<th>ADJ-Noun</th>
<th>ADV-ADJ</th>
<th>ADV-Verb</th>
<th>Noun-ADJ</th>
<th>Verb-Noun</th>
<th>Verb-ADJ</th>
<th>Verb-ADV</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numbers of Non-frequent ACL Items in CSC</td>
<td>526</td>
<td>35</td>
<td>23</td>
<td>16</td>
<td>4</td>
<td>19</td>
<td>3</td>
<td>626</td>
</tr>
<tr>
<td>Percentages of Non-frequent ACL Items in CSC</td>
<td>29.7%</td>
<td>28.2%</td>
<td>16.4%</td>
<td>25.8%</td>
<td>13.3%</td>
<td>6.1%</td>
<td>10.3%</td>
<td>25.4%</td>
</tr>
</tbody>
</table>
state/accomplish/define for purpose). Next, the two identified word groups are filtered with only the ones which “share” the most collocates with the target two words being left. This process results in a complete and manageable cluster. The words embedded in such clusters, finally, are further ranked in order of semantic relevance; that is, the more semantically related to the target words, the higher ranking (e.g. attain/accomplish for achieve and goal/objective for purpose). As the cluster involving achieve purpose suggests, an important function that collocation clusters are expected to perform is to detect/correct English learners’ collocation errors. Our system basically can show what pairs should be avoided (e.g. attain purpose) as well as what pairs are correct usages (e.g. achieve goal) which are semantically similar to the searched words.

In this study we applied the collocation-cluster techniques to ESP research and explored, for the ACL entries which were non-frequent in the CSC, whether corpus-derived clusters could automatically provide alternative usages. From the 626 non-frequent collocations, we tested the techniques on 60 randomly selected verb-noun or adjective-noun combinations. Those selected pairs were further checked or replaced by others to ensure that the 60 tested collocations did not include highly humanities- or social sciences-relevant ones (which apparently should not be expected to appear frequently in computer science texts). Each collocation then was fed into our system using the CSC as the main dataset to generate collocation clusters. The results collected were rather promising and encouraging. For the 60 collocations, collocation-cluster techniques successfully provided 50 semantically-similar alternative word pairs, showing a high rate of 83%. Some examples for the alternatives include construct argument for the ACL entry develop argument, demonstrate capability for demonstrate competence, primary concern for central concern, critical evaluation for critical examination, etc. These results again confirm Hyland & Tse’s (2007) viewpoint that people in different domains tend to have preferred expressions. If computer science students rely only on the ACL to learn academic collocations, it is likely that they will use several non-domain-specific academic pairs (e.g. profound effect, 0 times in CSC) instead of the ones that their colleagues prefer to produce in written academic texts (e.g. significant effect, 238 times).

Conclusion

The current research analysed the distribution of ACL entries in computer science research articles and reported an empirical study in which we successfully utilized collocation-cluster techniques to collect domain-specific academic collocations. The two types of results taken together suggest an innovative approach to gather academic collocations for a particular domain: ESP instructors can first examine which ACL entries are frequent and which are non-frequent in a domain-specific corpus and, for the latter, make use of corpus-derived collocation clusters to acquire semantically-similar alternatives. We in our presentation also discuss future improvements of collocation-cluster techniques, which we expect to make the techniques even more useful for both corpus linguists and ESP practitioners.

References


Bad language revisited: swearing in the Spoken BNC2014
Robbie Love (Lancaster University, UK)

1. Introduction

This paper reports on the use of McEnery’s (2005) approach to analyzing swearing in spoken British English to investigate the use of bad language words (BLWs) in a sample of the Spoken British National Corpus 2014 (Spoken BNC2014S) (Love et al. 2017 fc), comparing this corpus with the original Spoken British National Corpus (Spoken BNC1994) (Leech 1993). The Spoken BNC2014 comprises transcripts of spontaneous, present day, informal conversations among speakers of British English recorded between 2012 and 2016.

2. McEnery’s approach to swearing, and other typologies

McEnery (2005) analyses so-called bad language words (BLWs) in the Spoken BNC1994DS (demographically-sampled component). He uses the term swearing broadly to encompass a set of bad language words (BLWs), which includes literal and non-literal use of swear words (e.g. SHIT, FUCK) as well as other words which may be used offensively but which would not be considered swear words per se (e.g. PIG, TART). In addition to quantitative analysis of their distribution across the sociolinguistic categories of gender, age and socio-economic status, McEnery conducts qualitative analysis of each BLW using a bespoke bad language categorization scheme (originally developed for the Lancaster Corpus of Abuse – LCA, McEnery et al. 1999, 2000). He finds that “the use or lack of use of BLWs is a fault line along which age, sex and social class may be differentiated” (p. 50).

The LCA annotation scheme survives not without criticism. Ljung (2011: 12) defines bad language via a typology that has much in common with McEnery’s scheme but, crucially, excludes literal uses of swear words: “taboo words with literal meaning cannot be regarded as swearing”. Furthermore Ljung (p. 28) criticises several of the categories of the LCA scheme including Idiomatic set phrase, Imagery based on literal meaning and Pronominal form with undefined referent, suggesting that the scheme ought to be used with caution.

3. Method

The aim of this work is to replicate McEnery (2005) by analyzing a large set of BLWs in the Spoken BNC2014S and comparing their frequency, sociolinguistic distribution and use to that of the Spoken BNC1994DS, commenting on any changes in bad language over the last twenty years (for the sake of comparability with research on bad language in the BNC1994DS, I adopt McEnery’s definition of BLWs as opposed to that of Ljung, 2011). To do this, studying only the set of 50 BLWs from McEnery (2005) would be insufficient; I had to take into account the possibility of new BLWs having emerged since the early 1990s. I therefore extended the original list of BLWs by adding those that were included by the UK’s Office of Communications (Ofcom) in their guide to offensive language in broadcast media (Ipsos MORI 2016), as well as a set developed by Lutsky & Kehoe (2015). The use of these sources as bases for extending McEnery’s original list resulted in a new set of 173 BLWs.

This study was thus conducted as per the following:
(1) Search in the Spoken BNC1994DS and Spoken BNC2014S for each BLW in turn;
(2) Observe BLWs which have a frequency of zero in both corpora and eliminate from further analysis;
(3) Analyse BLWs which have changed in relative frequency the most drastically between the two corpora;
(4) Demographic distribution: select some of the most commonly occurring BLWs in both corpora, and record their frequency per speaker metadata category according to gender, age and socio-economic status;
(5) Annotate them according to the LCA annotation scheme (McEnery 2005);
(6) Comment on possible language change in light of differences between the two corpora in the BLWs’ speaker metadata distribution and categorization.

4. Data

Both corpora were accessed via Lancaster University’s CQPweb server (Hardie 2012). The Spoken BNC1994DS contains 5,014,655 tokens across 153 texts, while the Spoken BNC2014S contains 4,789,185 tokens across 567 texts.

In terms of corpus comparability, it could be argued that since neither of the Spoken BNCs were sampled with the explicit aim of studying BLWs, it is difficult to claim that the sampling conditions allowed for a comparable amount of BLW use. However, it can firstly be assumed that the Spoken BNC1994 facilitated the natural occurrence of BLWs given its surreptitious approach to recording (Crowdy 1993: 260). Secondly, the aim of the Spoken BNC2014 team was to facilitate the recording of conversations in a way which minimized intrusiveness beyond what was required of modern ethics procedures (Love et al. 2017). Harry Strawson, a Spoken BNC2014 contributor who submitted over a dozen recordings, claimed that “it was surprising how quickly people seemed to forget they were being recorded” (Strawson 2017).

5. Initial results: wholesale frequency analysis, and the case of FUCK

32 BLWs, including BUKKAKE, FATASS and PUNANI, were found to have a frequency of zero in both corpora. Many of these are described by Ofcom (2016) as having “low recognition” among focus group participants, and several were labelled as having been identified by less than 40% of participants in an online survey of the words. Based on this it is perhaps unsurprising that they do not occur in the corpora.

141 remaining BLWs occur at least once in either of the corpora. Among these, some were found to have decreased in relative frequency significantly (p<0.0001) between the 1990s and 2010s, including SPASTIC, CUNT and BUGGER. Those which have increased significantly (p<0.0001) include RETARD, DYKE and SHIT, while there are some – including ARSE, BITCH and FUCK – which have very similar relative frequencies in both corpora and have therefore shown stability.

Initially, I then looked at one of these stable BLWs – FUCK – in more detail. By applying the relevant steps as outlined in the Method, I could assess how FUCK has changed in spoken British English in the last two decades, according to the sociolinguistic variables of gender, age and socio-economic status.

The results suggest that in present-day spoken British English:

• FUCK is now used equally as frequently by male and female speakers.
• The use of FUCK peaks among speakers in their twenties and decreases with age, apart from the 60-69 group which has a higher frequency than 50-59.
The distribution of FUCK according to social class is similar to that of the Spoken BNC1994DS but only if the same classification scheme (Social Grade) is used. If a newer scheme is used (NS-SEC), then it is speakers in the middle of the scale that seem to use FUCK the most rather than those towards the bottom.

Wholesale frequency comparisons aside, the LCA annotation scheme revealed an interesting difference in the use of FUCK between the two corpora (Table 1):

Table 1. Annotation of FUCK in the Spoken BNC1994DS and Spoken BNC2014S using the LCA annotation scheme: most populated categories.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>E</td>
<td>Emphatic adverb/adjective: ‘He fucking did it’ ‘in the fucking car’</td>
</tr>
<tr>
<td>2</td>
<td>N</td>
<td>Premodifying intensifying negative adjective: ‘the fucking idiot’</td>
</tr>
<tr>
<td>3</td>
<td>G</td>
<td>General expletive ‘(Oh) Fuck!’</td>
</tr>
<tr>
<td>4</td>
<td>I</td>
<td>Idiomatic ‘set phrase’: ‘fuck all’ ‘give a fuck’</td>
</tr>
<tr>
<td>5</td>
<td>D</td>
<td>Destinational usage: ‘Fuck off!’ ‘He fucked off’</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rank</th>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>I</td>
<td>Idiomatic ‘set phrase’: ‘fuck all’ ‘give a fuck’</td>
</tr>
<tr>
<td>2</td>
<td>G</td>
<td>General expletive ‘(Oh) Fuck!’</td>
</tr>
<tr>
<td>3</td>
<td>F</td>
<td>Figurative extension of literal meaning: ‘to fuck about’</td>
</tr>
<tr>
<td>4</td>
<td>D</td>
<td>Destinational usage: ‘Fuck off!’ ‘He fucked off’</td>
</tr>
<tr>
<td>5</td>
<td>A</td>
<td>Predicative negative adjective: ‘Is it fucked?’</td>
</tr>
</tbody>
</table>

There appears to have been a shift in the way in which FUCK is used in British English conversation. The direction of change appears to be towards further generalization (and perhaps weakening) of this BLW, with idioms such as give a fuck, what the fuck and for fuck’s sake occurring very frequently and accounting for 31% of all instances. Furthermore there has been a rise in figurative extensions of the original meaning, including fuck it up, fuck me off and fuck around (14% of all instances). In turn, emphatic adverb/adjective forms of fuck, which did account for 55% of instances in the Spoken BNC1994DS, now only account for 2.7%. It seems then that FUCK has moved away from a modifying function which emphasizes other lexical words, and occurs much more idiomatically and figuratively than previously.

6. Conclusion

The design of the speaker metadata categories in the Spoken BNC2014 makes the new data comparable to the Spoken BNC1994 for the purposes of sociolinguistic analysis, and the case of FUCK as an example (along with others to be given in the presentation) suggests a clear change in use between the 1990s and 2010s, in addition to wholesale change and stasis of BLWs in the two corpora.

The compilation of the Spoken BNC2014 has facilitated large-scale diachronic analyses of spoken data on a scale which has until now not been possible. This study therefore exemplifies new opportunities – and challenges – in the sociolinguistic study of spoken data.
Acknowledgements

The research presented in this paper was supported by the ESRC Centre for Corpus Approaches to Social Science, ESRC grant reference ES/K002155/1.

References


The processing of collocations by native and non-native speakers of English: Evidence from ERP studies
Jennifer Hughes (Lancaster University, UK)

In this presentation, I present the results of two experiments which combine methods from corpus linguistics and psycholinguistics in order to find out how the brain responds to encountering adjective-noun bigrams which form strong collocations (e.g. clinical trials) compared to adjective-noun bigrams which are semantically plausible but do not form collocations (e.g. clinical devices). In Experiment 1, the stimuli were presented to 16 native speakers of English; in Experiment 2, the same stimuli were presented to 16 non-native speakers of English (all native speakers of Mandarin Chinese).

The 15 collocational bigrams used in these experiments were selected because of their high forward transition probability. This is calculated by dividing the number of times the bigram X-then-Y occurs in the written BNC by the number of times X occurs in the written BNC altogether (McEnery and Hardie 2012:195). For each collocational bigram, I manufactured a non-collocational bigram that was absent from the BNC. The nouns in both conditions were matched for frequency and length, and all bigrams were embedded into corpus-derived sentences. These sentences were edited to ensure that they were semantically coherent as standalone units (Spötter and McCarthy 2004:197) and that the sentence fragment preceding the experimental bigram was identical in both conditions. I also ensured that the preceding contexts created an equally “low contextual constraint” for the bigrams in each condition (Millar 2010:108).

The sentences were presented to participants at a rate of 500 ms per word (including a 200 ms interstimulus interval). Each participant wore a headcap containing 64 scalp electrodes, and these electrodes recorded some of the electrical activity of the brain while the participant was reading the stimuli. This method of “measuring electrical potentials in the brain by placing electrodes across the scalp” (Harley 2008) is known as EEG, or electroencephalography. One particular type of EEG study involves the measurement of ERPs, or event-related potentials, which are defined as “the momentary changes in electrical activity of the brain when a particular stimulus is presented to a person” (Ashcraft and Radvansky 2010:61). The ERP technique was used in the experiments discussed in this presentation, where the brain response was time-locked to the second word of each experimental bigram. These were then averaged across trials to isolate the experimental effect.

Previous studies have used the ERP technique to investigate semantic and syntactic processing. For instance, Kutas and Hillyard (1980) discovered that reading a semantic error elicits an N400, i.e. a negative voltage deflection occurring 400 ms after the onset of the error. Moreover, Osterhout and Holcomb (1992) found that reading a syntactic error elicits a P600, i.e. a positive voltage deflection that occurs 600 ms after the onset of the error. However, very few studies have used the ERP technique to investigate the processing of collocational errors, and those that have used ERPs for this purpose have tended to focus on idioms or other multi-word expressions. For instance, Molinaro and Cerrairas (2010:179-180) explicitly state that they focus on “idioms or clichés”, and Molinaro et al. (2013:124) state that their stimuli is comprised of “multi-word expressions”. The experiments discussed in this presentation therefore provide a unique contribution to the study of collocation by focusing on collocations that do not have clear beginning and end points and are not necessarily stored holistically.

The results of these experiments show that, for both native speakers and non-native speakers of English, reading the second word of a non-collocational bigram elicits an N400. It could be argued that this N400 is functionally distinct from the N400 discovered by Kutas and Hillyard (1980). While the N400 elicited in response to reading semantic errors typically has a
central-parietal scalp distribution, the results of my experiments show that the N400 elicited in response to reading collocational errors has an anterior-central scalp distribution. This replicates the results of my pilot study which was conducted on a different group of native English speakers.

From these experiments I can conclude that, for both native and non-native speakers of English, reading non-collocational bigrams elicits a quantitively distinct brain response compared to reading collocational bigrams. Interestingly, my results show that the N400 response is even larger for the non-native speakers. This finding is discussed in relation to the levels of prior exposure that both participant groups might have had to the bigrams outside the context of the experiments. Moreover, I discuss how the findings from these ERP studies provide evidence in support of a network model of language processing, whereby collocations are represented in the brain as transitions across a network.

References

We present a tool for close reading of texts that improves the intuitive understanding of corpus-based frequency measures (e.g., probability in context) using colour and font size.

Visualization techniques facilitate the understanding of data and their application in the digital humanities is actively researched. In their survey, Jänicke, Franzini, Cheema, and Scheuermann (2016) give an overview of visualization tools and the features used. Font size is rarely used as a feature. Word clouds, for example, scale words based on frequency, but they do not maintain sentence structure. Prism (Walsh, Maiers, Nally, & Boggs, 2014) uses size for the visualization of annotation agreement, with frequently annotated words shown in a larger font.

Corpus analysis tools like the Corpus Workbench (Evert & Hardie, 2011) and CQPWeb (Hardie, 2012) are useful for querying corpora and finding interesting text instances. Subsequent interpretation, however, is often difficult. In particular, corpus positions that are numerically annotated (see Table 1, third column) can be challenging.

<table>
<thead>
<tr>
<th>Word</th>
<th>POS</th>
<th>Bits</th>
</tr>
</thead>
<tbody>
<tr>
<td>It</td>
<td>PP</td>
<td>5.55</td>
</tr>
<tr>
<td>Grows</td>
<td>VVZ</td>
<td>8.26</td>
</tr>
<tr>
<td>In</td>
<td>IN</td>
<td>1.67</td>
</tr>
<tr>
<td>dry</td>
<td>JJ</td>
<td>4.95</td>
</tr>
<tr>
<td>ground</td>
<td>NN</td>
<td>3.28</td>
</tr>
<tr>
<td>.</td>
<td>SENT</td>
<td>3.81</td>
</tr>
</tbody>
</table>

Table 1. Sentence with part-of-speech and surprisal annotation.

In our work, we study the effects of surprisal of words in context on the sentence and text levels. Surprisal of words is defined as the negative logarithm of their probability and is measured in bits (cf. Genzel & Charniak, 2002). To this end we developed an interface for close reading, which uses font size and colour for the visualization of surprisal. The tool is implemented based on D3 (Data-Driven Documents; Bostock, Ogievetsky, & Heer, 2011) and can be used either stand-alone or for the visualization of surprisal in text instances discovered using CQP.

Figure 1 shows three sentences from a historical corpus visualized with our tool. The individual words are scaled based on surprisal. Pointing the mouse at the last occurrence of Arsenick, all of its occurrences in the text are highlighted by underlining. In addition, a pop-up window shows further information about the word. From the pop-up, one can see that the current word has a value of 19.24 bits and that Arsenick occurs six times in total (N=6). The values of all occurrences of Arsenick are in the range between 13.9 and 20.84 bits, the average being 19.15 bits. As a further help, words with low surprisal (e.g., function words) can be shown in a less intensive colour based on a threshold.
In summary, our tool can be used to spot outliers (instances with very high/low frequencies), which might be worth investigating further (e.g., in a concordance). The pop-up can provide further information about the variability of a word. Finally, one can recognize trends and distributional patterns that are not obvious from a tabular display as in Table 1.

References


‘This is sort of what you asked, you know’: A comparative corpus-based analysis of pragmatic markers acquired in a study abroad context

Giovani Santos (Mary Immaculate College, Ireland)

Studies on spoken language have greatly benefited from Corpus Linguistics over the recent years (Caines et al. 2016). One of the many insights from spoken corpora studies is that of the interactional and relational nature of spoken language in use (O’Keeffe et al. 2007, p.159). Indeed, face-to-face communication requires speakers to make use of strategies to convey their messages and orientate themselves through conversation in real time. Among such strategies is found a body of language employed to maintain the relationship between speakers, including pragmatic markers (PMs), which O’Keeffe et al. (ibid.) refer to as relational language.

This poster presents the reflections and results of a pilot study, part of a broader innovative PhD research project on second language development - with a focus on the acquisition and use of PMs by L2-users of English in a study-abroad context, viz. Brazilian university students living in Ireland. The data comprises two 30-minute informal interactions (unstructured interviews) between 13 participants and the researcher – one in English and one in Brazilian Portuguese – facilitating the construction of a bilingual corpus. Hitherto, 11 interactions have been recorded with 13 participants, of which 10 are dyadic and 1 is multi-party. The data amount to approximately 150,000 words across both languages.

Drawing on the English component of this bilingual L2 corpus and the one million word spoken corpus, the Limerick Corpus of Irish English (Farr et al. 2004), it will be possible to compare and contrast the interpersonal functions and procedural meanings of the most salient PMs found in the participants’ L2 against those of the Irish English data. The questions to be answered in this study are:

1. Have the Brazilian study-abroad participants acquired PMs found in Irish English?
2. Is there a difference in the use of pragmatic markers between the L1 and L2 corpora?
3. Are the L2-users pragmatically successful in their use of PMs?

While Irish English has been much studied, this is the first study to focus on the influence of this language variant on a sub-group of L2-users. What is more, although this study compares and contrasts the functions and meanings of PMs between L2- and L1-users of English, it does not view the L2-user within a deficit model and will, instead, focus primarily on pragmatic competence of spoken language in use in the context of interlanguage. In that regard, this study concurs with the view of Prodromou (2005) that L2 production merits analysis in its own right, rather than applying a ‘native-centric’ approach.

Although the data sample selected for this pilot study may be considered small, it is argued that it can allow for a more detailed qualitative analysis of
the spoken material produced by this sub-group of L2-users of English, throwing light on important linguistic features that, otherwise, could go unnoticed if analysed from a larger corpus. As Koester (2010, p.67) notes, “a relatively modest corpus may still yield robust and powerful findings”.

A future step in this PhD project will be the addition of the Brazilian Portuguese component of the L2 corpus to the analysis. In working through this bilingual L2 corpus environment (Granger 2002), it is expected that it will be possible to illuminate features of the participants’ L2 based on their cultural and social immersion in a native English speaking context, as well as on their L1 transferences.

References


Metaphorical Constructions in Modern Economic Discourse: A Large Scale Corpus Analysis
Jenny Lederer, Helena Laranetto and and Guy Brown
(San Francisco State University, USA)

The idiosyncratic nature of lexicalization patterns is fact of all languages and raises interesting questions concerning speakers’ linguistic knowledge. Our study focuses on one aspect of idiomacity –how metaphors are revealed in English talk and text. Metaphorical patterns can be identified, to some extent, by probing native speakers about their own language use. But this technique often fails because speakers have inconsistent explanations of what they can and do say. Recently, corpus linguists have begun to use corpus frequency statistics to investigate the relationship between conceptual metaphor, grammatical constructions, and the set phraseology of topical discourse (Demmen et al., 2015; Koller et. al., 2008; Lederer, 2016). The conceptual metaphors we use to understand and talk about abstract ideas like the economy and a myriad of other ‘target domains’ –education, career, life, and love –are primarily unconscious (Lakoff & Johnson 1980); therefore, the best way to probe nuanced patterns of how these abstract ideas are lexicalized, or encoded in words and phrases, is to look at sizable collections of data.

Economic language, specifically, is based on a complex conceptual model involving several robust metaphors including MONEY IS LIQUID, THE ECONOMY IS A SHIP; THE ECONOMY IS A WEATHER EVENT; and the ECONOMY IS AN AILING BODY (Boers and Demecheleer, 1997; Shenker-Osorio, 2012). These metaphors work together to produce a set of interrelated metaphorical tokens. Source triggers combine with target triggers in close proximity, comprising set phraseology typical of economic discourse. When it comes to metaphorical collocation patterns, related terms show unequal frequency patterns. For example, when discussing business and finance, speakers use cash flow more frequently than cash stream, but when discussing investment, speakers use income stream and not income flow (Lederer, 2016, p. 545). Flow and stream mean the same thing, more or less, but they don’t equally participate in economic jargon. Frequency differentials tell us something about the interaction between the grammar of language and our conceptual system (Musolff, 2006). In this specific case, they illustrate interesting properties of how money is metaphorically understood as liquid –cash flows around the world relatively unimpeded. Income, however, is assumed to exhibit unidirectional movement –from investment or job to investor or worker. Thus, the movement of income is more like a stream –flowing from origin to destination in one direction only.

Drawing on patterns like these, we use seed words like financial and storm (identified as frequent metaphor triggers in The Economist magazine) to probe fixed and semi-fixed economic phraseology. Using the Corpus of Contemporary American English (Davies, 2008–), our large-scale examination includes 12 frequent metaphorical target triggers (bank, invest, market, firm, debt, finance, rate, price, capital, growth, economy, money) combined with 84 source triggers (including lemmas like sail, storm, channel, conditions, freeze, cloudy, turbulent, ailing, and health) to produce 1008 pairings, i.e. investment freeze and turbulent market. Corpus methodology not only allows for a close examination of how conceptual metaphors are encoded in lexis, but also an intimate examination of how conceptual metaphors populate syntactic constructions (cf. Sullivan 2009; 2013). Thus, each lexical combination was also evaluated for its rate of occurrence in five syntactic constructions: NP-of-NP (captain of finance; wave of capital), A-NP
(economic headwinds; ailing economy), N-N compound (debt bubble; bubble market) V-NP (steer the economy), and X is Y (the economy is a sinking ship).

All 2,016 ordered combinations (e.g. invest* flood*/ flood* invest*) were searched thoroughly by allowing for up to five intervening wild cards (i.e. flood of American investment) in order to catch most, if not all, instances of relevant tokens (totaling approximately 13,000 manual searches in COCA). Metaphorical tokens were identified by hand and added to our database –coded for metaphor, specific trigger, and syntactic construction. Coding for metaphorical use followed the MIPVU method, a commonly accepted standard for metaphor identification (Pragglejaz Group 2007; Steen et. al. 2010). This method, applied to all searched collocations, involves 1) reading each concordance line corresponding to collocation searches, 2) establishing the contextual meaning of the collocation in question, and 3) marking the phrase as metaphorical when it is not used to signal a more basic, literal referent. For example, a floating market would not be collected as a metaphorical phrase when it references a farmers market in Southeast Asia built on floating platforms; however, a floating market, used to reference a numerical index of a particular economic market, would be marked as metaphorical.

Our findings from COCA show that asymmetries in source-target pairings abound in metaphorical economic phraseology. Acknowledging metaphorical mappings are more constrained by autonomous target domains, Sullivan (2013) hypothesizes a preference for target-domain language to populate semantically more autonomous syntactic positions in modifier-noun constructions, predicting that source and target triggers should align with predicate-specifying and argument-specifying syntactic positions, in that order. Through detailed type and token counts, our results confirm that source domains function as predicational material used to structure the target domain, explained by asymmetrical conceptual dependencies built into the syntax of simple constructions (cf. Dancyngier & Sweetser 2014; Langacker 1991). Given a lexeme’s origin –source or target- we can predict syntactic alignment when used in metaphorical phraseology, market climate is metaphorical, climate market is not. Exceptions to these strong tendencies are explained through genre-specific lexicalization processes in which predicational terms like bubble (market bubble) establish themselves as domain modifiers (bubble market) in economic jargon.

In addition to reporting syntactic patterns of metaphor instantiation, we present findings on three other parameters of lexicalization. One way to judge or measure the lexicalization patterns of conceptual metaphors is to consider conventionality as a scale of idiomaticity (Gibbs 1984). Frequency statistics can be seen as a proxy for conventionalization. On one end, highly frequent collocations of set source and target triggers occur as common fixed or semi-fixed phrases; on the other end, more creative and novel pairings will be less frequent in corpus data, and, thus, not considered conventionalized, metaphorical jargon. For example, economic recovery, economic conditions, market conditions, economic climate, and financial health, together produce over 3,000 tokens in our 8506 token sample; whereas others pairings, with similar meanings like economic sickness or economy forecast, occur only once each.

We also present results from a second line of investigation examining distributional coverage: which source triggers are the most permissive in their ability to co-occur with target triggers and vice-versa? Borrowing from Hausmann (2007), we refer to this measure as ‘metaphorical combinability’ and report measurements for each trigger in the sample. We explore factors that might contribute to the high combinability scores for certain source triggers like flow*, freeze*, and recover*. Combinability seems to correlate with a lemma’s rate of lexical variation, base rate frequency in economic discourse, and
keyness score—a lemma’s relative frequency in economic discourse compared to overall use in English (Cheng 2012; Ahmad 2005).

Finally, we investigate unequal frequencies in specific source and target pairings. We explore why there is a token frequency difference between sequences of the same triggers when both sequences can be used metaphorically. For example, as a domain-modifier construction (A-N) there are 270 occurrences of financial health but only two instances of the predicate-modifier construction, healthy finances. Both possibilities come from the same metaphor, THE ECONOMY IS AN AILING BODY; however, the source and target triggers do not fill the A-N slots at the same rate. Based on the examination of base-rate scores and keyness statistics, we believe these asymmetries have less to do with overall word frequency differentials and more to do with discourse context: public discussion of economic topics focuses more on macro- versus micro-economics.

References


The use of ‘will’ in present tense narrative
Reiko Ikeo (Senshu University, Japan)

Over the past decade, more writers of English have begun to employ the present tense as the primary mode for their narratives instead of the past tense, which has up until recently been regarded as the norm for narrative tense. One example which illustrates this current trend is that out of 102 novels which were shortlisted for the Man Booker Prize in the UK between 2000 and 2016, 31 of them, that is 30.4%, narrate in the present tense.

The choice of tense in fiction is not a mere syntactic operation but a stylistic and narratological decision. Genette (1980) explains the function of tense in narrative as that ‘in which the relationship between the time of the story and the time of the discourse is expressed’ (p.29). When this narratological perspective is applied to each tense, the past tense in a narrative signifies that the narrated events occurred prior to the act of narration. On the other hand, in fiction in which the present tense is used as the narrative mode throughout, this time lag between the occurrences of events and the speaker’s enunciation is supposed to disappear. This simultaneity of narration and events seems to be the aspect that contemporary present tense narratives exploit and benefit from most. The use of the present tense, however, is not restricted to expressing the narrator’s or a character’s here-and-now. In fact, as Fludernik (2003: 201) and Huber (2016: 13-14) point out, literary usage of the present tense does not necessarily correspond with the time being narrated. As well as laying out the situation of narration, the present tense can be used to portray past events and a character’s memories or to depict a character’s inner monologue.

The present tense in narrative not only signifies a simultaneity of action and narration but also implies a state that is open to the unknown future, where neither the narrator nor the character knows the consequences of the present action. A keyword analysis, using two small corpora reveals an aspect which can stylistically distinguish present tense narrative from past tense narrative. One corpus which I used for this study was reproduced from the serious fiction section of the Lancaster Speech, Writing and Thought Presentation corpus (Semino and Short 2004). Out of the twenty different fictional texts constituting this section in the original corpus, 1st-person narratives and present tense narratives were excluded. The remaining ten texts, each of which was comprised of 2,000 words, were used for a smaller corpus of past tense narrative (the past-tense corpus). A new corpus was compiled from 2,000 word texts of ten novels written in the present tense published after the year 2000 (the present-tense corpus). The narratives in both corpora all have 3rd-person narrators instead of 1st-person or 2nd-person narrators.

Two keyword lists were created based on these two corpora by comparing the two word lists, each of which was created from one of the two corpora, using the corpus concordancer Wordsmith Tools (Scott 2016). One of the pronounced features of the lists is the high keyness value of the modal verbs ‘will’ in the present-tense corpus and ‘could’ in the past-tense corpus. Other verbs on the keyword lists tend to correspond to the tense feature of each text: the keywords in the present-tense corpus compared with the past-tense corpus include ‘is’, ‘says’, ‘has’, ‘does’ and ‘thinks’ while
the keywords in the past-tense corpus compared with the present-tense corpus include ‘was’, ‘had’, ‘said’, ‘were’, ‘looked’.

It can easily be assumed that the modal verb ‘will’ will appear in a character’s direct speech/thoughts in the narratives of both tenses. It does in both corpora. However, the number of occurrences in the past-tense corpus is very small, and the occurrences are limited to particular texts. On the other hand, in the present-tense corpus ‘will’ occurs in all the texts except one and not only in characters’ direct speech and thought presentation but also in narration and characters’ indirect discourse presentation.

The concordance lines containing ‘will’ show the stylistic peculiarity of present tense narrative which allows the use of ‘will’ in narration and indirect forms of discourse presentation. Particularly in narration, the modal verb ‘will’ seems to reflect the viewpoints of both a character and the narrator to whom the consequences of the present events or actions are unknown. One example which illustrates this stylistic function in present-tense narrative is from Hilary Mantel’s *Bring Up the Bodies*:

(1) ‘Well, you know,’ he (Thomas Cromwell) says. ‘We can’t go throwing ambassadors out. Because then we don’t get to know anything at all.’

Truth is, he is not afraid of Katherine’s intrigues: the mood between France and the Empire is at the moment unremittingly hostile, and if open war breaks out, the Emperor *will* have no troops to spare for invading England.

(Mantel 2012: 76, my emphasis)

If the passage above is rewritten in the past tense, the same degree of intensity and incompleteness due to the unknown future state which is implied by the modal verb might not be maintained by the substituted ‘would’.

Although the use of modal verb ‘will’ in narration and indirect forms of discourse presentation is grammatically possible only in present tense narrative, the tense system needs to be considered separately from the modal system in analysis as Palmer (1990) argues. ‘Will’ belongs to the modal system, not the tense system, which is morphologically marked as present or past. The modal verb ‘will’ tends to be used where there is reference to a general envisaged, planned, intended or hoped for state of affairs, as opposed to a statement that a specific event will in fact take place (Palmer 1990: 140). In the example above, the situation in which the Emperor cannot afford troops for invading England is envisaged by Cromwell (and the narrator), not forecasted neutrally as a future state. Futurity implied by ‘will’ derives from modality rather than tense.

Another modal verb ‘could’ is found on the keyword list of the past-tense corpus. The concordance lines of ‘could’ in the past-tense corpus show that approximately 72.7% of the occurrences of ‘could’ are associated with the past tense especially when it expresses the ability of the subject or the possibility of the event discussed. On the other hand, about 27.3% of the occurrences function as a means of indirectness, or as a part of conditionals and metaphorical expressions. This suggests that such usage of ‘could’ is related more to pragmatic politeness and rhetorical effects than to the past tense of the narrative.
References


Developing a TV drama corpus-informed general spoken formulas list for elementary-level EFL learners

Hiroya Tanaka (Hokkai-Gakuen University, Japan)

This study aims to develop a general formulaic sequences list for elementary-level EFL learners informed by an American TV drama corpus. The use of formulaic language is considered to play a significant role in spoken discourse as well as in written discourse (e.g., Boers, Eyckmans, Kappel, Stengers, & Demecheleer, 2008) and researchers have developed lists of formulaic languages for various multiword units such as phrasal verbs (Garnier & Schmitt, 2015), written and spoken academic formulas (Simpson-Vlach & Ellis, 2010), and non-transparent multiword expressions (Martinez & Schmitt, 2011). There has been, however, little attention paid to date to pedagogical formulaic-sequence materials for informal conversations which could be used for elementary-level EFL learners, although oral communication is largely informal in nature even in workplaces (e.g., Crosling & Ward, 2002).

Davies (2012) claims that a TV drama corpus, the Corpus of American Soap Operas in his case, is at least spoken-like and contains more words dealing with everyday life and personal relationships than spoken proportion of BNC and COCA. Thus, it is fair to say that a TV drama corpus has a potential to inform the use of formulaic sequences in informal conversations. The author has developed a corpus of American TV sitcom scripts to reveal the use of formulaic sequences in informal conversations and to make a pedagogical general spoken formulas list for elementary-level EFL learners. The scripts were collected from various sitcom fansites, which resulted in a five-million word corpus of 18 sitcom titles from 2006 to 2016. Following some of the previous studies (Coulmas, 1979; Wood, 2010; Wray & Namba, 2003), the author created and used a checklist in order to identify spoken formulas in the corpus. The requirements in the checklist are; each formula has to 1) have two or more words (2- to 5-gram), 2) have the frequency of more than 10 per million words, 3) have a pragmatic integrity, 4) be transparent rather than opaque. The first two requirements were checked using a corpus concordance software, CasualConc (Imao, 2016), and the last two requirements were checked manually using learner dictionaries on the Internet. The poster will present a tentative list of general formulas informed by the five-million word American sitcom corpus and will discuss further possibilities of the list.

References


Do English textbooks reflect the actual use of English?: the present perfect and temporal adverbials
Kazuko Fujimoto (Soka University, Japan)

1. Introduction

Fujimoto (2003) states that Japanese university and high school students have a tendency to depend on temporal adverbials in their learning and using the present perfect and that they seem to have difficulty in using the present perfect without the help of typical temporal adverbials which textbooks and grammar books provide. According to Schlüter’s (2006) comparison of major corpus-based studies, less than 50% (between 29% and 45%) of the present perfect instances are used with temporal adverbials (p. 143), and his findings were confirmed by Hundt & Smith (2009). In the latter study, less than 35% of the present perfect instances co-occur with temporal adverbials (ranging between 32% and 34%) (p. 63). McGrath’s (2013) comment, “the textbook . . . influences what teachers teach and what and to some extent how learners learn” (p. xii), motivated my corpus-based analysis of textbooks. The aim of this study is to examine whether English textbooks published in Japan for high school students provide the real use of the present perfect and temporal adverbials, comparing a textbook corpus with native speakers’ corpora.

2. Methodology

Three types of corpora were used in this study. The first is a longitudinal learners’ corpus of about 100,000-word written English by 87 second-year Japanese university students who took an academic writing course in the department of the English language in 2009, 2010, and 2012 (Fujimoto Corpus [hereafter FC]). The students submitted their writing assignments every two weeks, eleven times in total per year. The average of the students’ TOEIC-IP scores is 458.9 (Range: 210-755; Median: 457.5; SD: 125.4). The second is four sub-corpora in the Brown family of corpora: Fiction and General prose sub-corpora in FLOB (British English [BrE], 1991) and Frown (American English [AmE], 1992) (FLOB_Fiction, FLOB_GP, Frown_Fiction, and Frown_GP). The size of these sub-corpora is about 300,000 words in FLOB_Fiction and Frown_Fiction and about 500,000 words in FLOB_GP and Frown_GP. The third is a textbook corpus of about 100,000 words, which is composed of written English data in passages, exercises, and other sections in six high school textbooks for the course English Communication I (English Communication I Corpus [ECIC]). The

---

1 Brown Family (C8 tags): powered by CQPweb was used. Available at http://cqpweb.lancs.ac.uk/.
2 The occupancy rate of these six textbooks has been more than 40% in total from 2013 to 2015.
reason why the textbook corpus was compiled from these textbooks for English Communication I is that this course is the only one required English course (and the other English courses are elective), and in this course, all the grammatical items (including the present perfect) specified in the government curriculum guidelines are to be taught. The average readability score (Flesch-Kincaid Reading Ease) of the textbooks is 82.6. First, FC was analyzed to see the university students’ use of the present perfect and temporal adverbials, and then ECIC and the four sub-corpora in FLOB and Frown were compared to examine the frequency of the present perfect and of the present perfect used with or without temporal adverbials. The data in FC and ECIC were analyzed with the computer software AntConc. The finite simple present perfect (SPP) was examined in this study (and the present perfect passive, the present perfect progressive, and the present perfect passive progressive were not included in the analysis). The instances of have [has] got whose meaning is “possess” and have [has] got to, and the present perfect forms containing modal idioms and semi-auxiliaries (e.g., have [has] been to, have [has] been able to) were manually excluded, based on Quirk et al. (1985, pp. 137-138, pp. 142-143).

3. Results and discussion

In FC, the raw frequency of the SPP is 174, and 64.4% of all the instances (112 instances) are used with temporal adverbials. 64.4% is much higher than the rate shown in Schlüter (2006) and Hundt & Smith (2009). The results of the analysis of the four sub-corpora (FLOB_Fiction, FLOB_GP, Frown_Fiction, and Frown_GP) are presented in Figure 1. In all the sub-corpora, the SPP without temporal adverbials is more common than that with temporal adverbials. Less than 40% of the SPPs are used with temporal adverbials. My analysis of the four sub-corpora provides further evidence supporting the previous studies by Schlüter (2006) and Hundt & Smith (2009): the SPP without temporal adverbials is much more frequent.

---

3 READABLE.IO. © 2011 - 2017 Readable.io. Available at https://readable.io/ (accessed March 11, 2017). The text with the score of 80.0 to 90.0 is considered to be easy to read.

The comparison of the frequency of the SPP in ECIC and in the four sub-corpora indicates that the SPP is more often used in ECIC than in FLOB_Fiction, Frown_Fiction, and Frown_GP. Log-likelihood tests show that the difference between ECIC and FLOB_Fiction and between ECIC and Frown_Fiction is significant at $p<0.0001$, and the difference between ECIC and Frown_GP, at $p<0.001$. There is no significant difference between ECIC and FLOB_GP. However, when passages in the six textbooks are compared with the four sub-corpora, no significant difference is observed between ECIC and FLOB_GP or between ECIC and Frown_GP, though the difference of frequency is significant between ECIC and FLOB_Fiction ($p<0.001$) and between ECIC and Frown_Fiction ($p<0.0001$). The results from my analysis indicate that overall, the SPP is more used in the textbook corpus than in the native speakers’ sub-corpora and that the textbook passages are closer to the BrE and AmE general prose than the BrE and AmE fiction in the use of the SPP. As for the frequency of the SPP co-occurring with temporal adverbials, 52.3% of all the SPPs in ECIC are used with temporal adverbials. Figure 2 shows the results of the analysis of each of the six textbooks. In the three textbooks *All Aboard!*, *COMET*, and *ELEMENT*, the frequency of the SPP with temporal adverbials is higher, while in the rest of the three textbooks *CROWN*, *VISTA*, and *Vivid*, the SPP without temporal adverbials is more common. In Table 1, the frequency of the SPP with temporal adverbials according to the sections in the textbooks (i.e., passages, exercises, and other) can be seen. In passages, the SPP with temporal adverbials is less used than that without temporal adverbials in all the textbooks but *VISTA* in which 50% of all the SPPs are used with temporal adverbials. It should also be noted that the SPP with temporal adverbials is most frequently used in exercises, which include grammar exercises, in all the textbooks except *VISTA*. It may be said that the SPP with temporal adverbials is likely to be overemphasized in grammar exercises in textbooks.
Table 1. SPPs Co-occurring with Temporal Adverbials

<table>
<thead>
<tr>
<th></th>
<th>All Aboard!</th>
<th>COMET</th>
<th>CROWN</th>
<th>ELEMENT</th>
<th>VISTA</th>
<th>Vivid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Passages</td>
<td>2</td>
<td>2</td>
<td>7</td>
<td>8</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>(8.0%)</td>
<td>(7.1%)</td>
<td>(22.6%)</td>
<td>(30.8%)</td>
<td>(50.0%)</td>
<td>(20.0%)</td>
</tr>
<tr>
<td>Exercises</td>
<td>22</td>
<td>26</td>
<td>22</td>
<td>16</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>(88.0%)</td>
<td>(92.9%)</td>
<td>(71.0%)</td>
<td>(61.5%)</td>
<td>(50.0%)</td>
<td>(80.0%)</td>
</tr>
<tr>
<td>Other</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(4.0%)</td>
<td>(0.0%)</td>
<td>(6.5%)</td>
<td>(7.7%)</td>
<td>(0.0%)</td>
<td>(0.0%)</td>
</tr>
<tr>
<td>Total</td>
<td>25</td>
<td>28</td>
<td>31</td>
<td>26</td>
<td>14</td>
<td>10</td>
</tr>
</tbody>
</table>

4. Conclusion

On the basis of my corpus analysis, the previous corpus research findings are not entirely reflected in Japanese English textbooks examined, and the grammar exercises in textbooks may have some influence on the students' use of the SPP with temporal adverbials. Tomlinson (2010, p. 89) explains gaps between what students are taught about English use and how English is actually used, pointing out the following as one of the main gaps: “Language use which has been captured by corpora but which is distorted for the sake of pedagogic simplification and convenience.” One interpretation of my findings might be that the co-occurrence of the present perfect with temporal adverbials in textbooks is unhelpfully exaggerated. However, given the non-existence of a dedicated present perfect marker in Japanese, I would argue that emphasizing co-occurrence with temporal adverbials may be a useful tool in the teaching of the English present perfect to Japanese students, especially to
less proficient students. It would also be needless to say that as students’ English proficiency increases, the actual usage of the present perfect with or without temporal markers should be more reflected in teaching materials.

References


Using the CORE Corpus and a multivariate analysis to (re)examine the impact of register and structural factors on that/zero complementizer variation in five mental state verbs
Christopher Shank (Bangor University, UK) and Koen Plevoets (FLAMES – Lstat, Belgium)

This paper examines the distribution of that/zero complementation alternation patterns, and related claims, concerning the proposed roles that register (specifically type of register) and concomitant structural factors/variables within the matrix and complement clauses play as predictors of the presence of the zero form, within five mental state verbs (MSVs) viz. think, believe, feel, know and understand.

I. I think that I want to have a baby. (Register: Narrative, CORE)
II. He believes that governments must not mandate prayer. (Register: Opinion, CORE)
III. They feel that their job is done. (Register: Personal Blog, CORE)
IV. She knows that it's not an option. (Register: Short Story, CORE)
V. Everyone here understands that it's not easy. (Register: Informational Description, CORE)

The methodological framework for this paper was developed around a comprehensive review of that/zero variation literature which identified thirteen different structural features that have been claimed to predict the presence of the zero complementizer. This same process was used to identify the role(s) that different types of register also apparently play in complementizer selection processes. This review resulted in the formulation of the following argument/claim: the more formal the register (and genre) the more likely the presence/use of the overt that complementizer form. To be more specific, authors have argued for a ‘register continuum’ whereby (and conversely) the greater the informality of both written and spoken texts (Bryant, 1962: Rissanen, 1991), the greater the degree of subjectivity in a text (Storms, 1966), the presence of a more casual style (Elness, 1984), the more informal the text type (spoken versus written, private versus public , unscripted/scripted) and the greater the informality of writing style (i.e. non-academic/academic and non-professional/professional) (Kaltenbock, 2004) are all claimed to act as predictive factors for the use/presence of the more 'informal' zero complementizer form. These structural and register based variables were then used to guide the study’s methodological design, data coding parameters and research questions.

Using the 50 million word Corpus of Online Registers of English (CORE) a total of 45,000 hits (for all 5 verbs) were randomly extracted from each of CORE’s 32 different register categories (Biber, Egbert & Davies, 2015). Once the data was extracted all of matrix plus complement that/zero constructions were coded for register category plus 28 structural variables including person, tense, polarity, and presence of modal auxiliaries, syntactic complexity, and complement clause subjects. Statistically sufficient sample sizes (n>40) for all register categories were extracted and a stepwise regression analysis is used to examine the statistical significance of 13 structural factors (as summarized in Kaltenbock 2004, described in Torres Cacoullos and Walker 2009 and presented in Author 1, Author 2 and Cuyckens, H, 2014) and register in regards to the selection of that/zero usage patterns for all five verbs in each of the 32 register categories.

The research questions guiding this paper are the following: (i) whether the conditioning factors proposed in the literature indeed predict the zero form, (ii) to what extent these factors interact, (iii) whether the predictive power of the conditioning factors...
becomes stronger or weaker depending upon the type of register vis-à-vis formality, (iv) the impact of register itself as a variable on predicting the presence zero form, (v) what is nature of the relationship (if any) between the each of the five MSPs, the register categories and that/zero variation in general – do any MSPs show a tendency to be utilized within particular registers in terms of either significant that or zero complementizer usage and (vi) what new or additional insights are gained with this approach compared to previous studies regarding the impact of register on that/zero variation patterns conducted with significantly smaller corpora containing substantially fewer register categories and non-multivariate modelling (Finegan & Biber, 1995; Kaltenbock 2004).

The results reveal varying degrees of significance for each of the 13 matrix and complement clause features, however; stronger significance and implications are revealed when additional variables (e.g. polarity, length of the subject, verb type, the effect of register(s) as a variable etc.) are incorporated via a ‘weighted’ variable analysis. These findings are used to identify the structural factors which are significant in predicting the presence of the zero complementizer form for each of the 5 verbs and within and across each of the 32 register categories. These findings will then be used to facilitate a discussion concerning the implications for using this type of statistically driven multifactorial approach.

References


‘Que(e)rying’ identity: Intersections of straight and gay male sexual identities in ‘casual encounters’ ads
Łukasz Woźniakowski (Adam Mickiewicz University in Poznań, Poland)

Sexuality can be regarded as an umbrella term which encompasses both sexual identity, that is, “the social framings through which individuals and groups are socially categorised (by themselves or others) based on their sexual orientation, beliefs about their sexuality, and/or their sexual practices” and sexual desire, understood as erotic desires and practices (Queen, 2014, p. 204). Some researchers contend that sexual identity is inextricably linked to sexual desire and that they both should be studied together as interconnected dimensions of sexuality. Among them, Bucholtz and Hall (2004) propose a view of sexual identity as the result of intersubjectively negotiated both practices and ideologies. This approach takes the view that sexual identity, like any other social identity, is inherently relational, unstable and contingent on a specific local context.

An interesting example of sexual identity as inherently relational and socially contingent is the discursive construction of ‘authentic’ heterosexuality in personal ads posted by self-identified straight men seeking same-sex sexual encounters (Ward, 2008). An analysis of this intersection of straight and gay sexual identity may give insights into the resources people draw on to construct themselves as inhabiting hetero- or homonormative worlds, as well as the role of other socially salient categories in the construction of different sexualities.

In this study, which is a work in progress, I will carry out corpus-assisted discourse analysis of personal ads posted by self-identified straight men seeking sex with men from the ‘casual encounters’ section of the bulletin board website craigslist.com. The aim is to explore the intersections of straight and gay male sexual identities and to identify the attitudes and ideologies drawn on by posters to construct their sexual identity. Also, I will look at how gay and straight sexualities intersect with other socio-politically salient categories such as race, age, gender or social class. Methodologically, this study draws on corpus analytic methods such as keywords, collocation and concordance analysis, and on Bucholtz and Hall’s (2004) tactics of intersubjectivity analytical model. By combining the two approaches, this study will provide not only some quantitative evidence of the discursive constructions of sexualities, but also some qualitative insights into the construction of straight and gay sexualities in the specific local context of online personal ads.

References

KeyWords of success – what words are associated with success in online Citizen Science?
Glenn Hadikin (University of Portsmouth, UK)

This paper aims to answer one primary research question: What words are associated with success in online Citizen Science? Citizen Science (CS) is a form of science crowdsourcing where members of the public come together but rather than give money they give their time to help various universities and other research institutions with their research. The concept arguably began in 1900 when members of the US Audubon society asked its members to count birds on Christmas day rather than participate in the tradition of shooting them (Silvertown 2009). Since then citizen scientists have been involved in helping researchers study the evolution of snails, recording air, soil and water quality and they have been conducting a survey of invasive species in the USA (Silvertown, 2009). In the last five years, there has been a burst of research activity into the phenomenon of CS, and, since 2009, the website Zooniverse.org has developed into the world’s largest CS site where volunteers join projects online and help label or classify data. It is timely for the research community to ask what influences these volunteers - especially the effect of their linguistic choices.

In this work-in-progress, techniques from corpus linguistics have been combined with methods from economics to compare the language used in seven weaker CS projects with six more successful ones from the Zooniverse. Measures of success are taken from Cox et al. (2015), and include various measures of public engagement and scientific success such as peer-reviewed publications, user posts and blog posts.

The data

<table>
<thead>
<tr>
<th>Zooniverse project</th>
<th>Public engagement score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Snapshot serengeti</td>
<td>0.614</td>
</tr>
<tr>
<td>Galaxy zoo old</td>
<td>0.453</td>
</tr>
<tr>
<td>Galaxy zoo new</td>
<td>0.416</td>
</tr>
<tr>
<td>Planet hunters</td>
<td>0.382</td>
</tr>
<tr>
<td>Planet four</td>
<td>0.25</td>
</tr>
<tr>
<td>Seafloor explorer</td>
<td>0.118</td>
</tr>
<tr>
<td>Ancient lives</td>
<td>0.071</td>
</tr>
<tr>
<td>Old weather</td>
<td>0.049</td>
</tr>
<tr>
<td>Milky way</td>
<td>0.043</td>
</tr>
<tr>
<td>Bat detective</td>
<td>0.041</td>
</tr>
<tr>
<td>Cyclone center</td>
<td>0.03</td>
</tr>
<tr>
<td>Solar stormwatch</td>
<td>0.027</td>
</tr>
<tr>
<td>Moon zoo</td>
<td>0.015</td>
</tr>
</tbody>
</table>

Figure 1: Public engagement score for 13 Zooniverse projects based on Cox et al. (2015)
Cox et al. (2015) use a positioning matrix to rank 18 Zooniverse projects based on 6 'contribution to science' elements and 6 'public engagement' elements. This paper makes use of their public engagement ranking criteria only. As an example, the collaboration element of public engagement is calculated as the following:

\[
\text{Number of papers with citizen scientist authors} \times \frac{1}{(\text{Project age})^2}
\]

Wright et al. (2016) built corpora of the text-based discussion forums associated with 43 Zooniverse projects. The 13 shown in figure one were both available as corpora and have been given a public engagement score in Cox et al. (2015) so were selected for this study.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Number of tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>topsix</td>
<td>5 851 392</td>
</tr>
<tr>
<td>bottomseven</td>
<td>1 220 710</td>
</tr>
</tbody>
</table>

Figure 2: Number of tokens in each of the two Zooniverse corpora

Figure two shows the size of the two corpora that were used for a KeyWord analysis using WordSmith tools 6.0 (Scott, 2012). The corpus 'topsix' is simply all the forum data from the six projects with the highest engagement scores and 'bottomseven' contains all the forum data from the seven with the lowest scores. Note there is different subject matter being discussed in each project forum so that must be considered when interpreting results. The risks associated with a small reference corpus highlighted in Berber-Sardinha (2004) are also acknowledged.

**KeyWords of success**

<table>
<thead>
<tr>
<th>KeyWord category</th>
<th>KeyWords</th>
</tr>
</thead>
<tbody>
<tr>
<td>space</td>
<td>UNIVERSE GALAXIES GALAXY STAR STELLAR STARDUST DARKNESS SPIRAL MERGER PLANET TRANSIT MERGE REDSHIFT PLANETS ASTEROID</td>
</tr>
<tr>
<td>religion</td>
<td>ANGELS HEAVENS GOD</td>
</tr>
<tr>
<td>grammar</td>
<td>JUST WE INTO ME RE THEN IS NOT WITHIN MY YOU ITS EITHER ALL WHERE AND A CANNOT SUCH I INFANT WHY SHE YOUR IM YA</td>
</tr>
<tr>
<td>names</td>
<td>PADDY IZZY JOHN EDD WEEZ WAVENEY TSERING KITHARODE MEG PLUK FIRESTORM ALICE ROY LIZ TERRANCE HANNY CURTIS JOHNSON DAVE COOK ANDREW SOPHIE HIGGS SUE PLANCK PETER PAUL ROGERS EINSTEIN ADAMS PAT ALBERT</td>
</tr>
<tr>
<td>everyday</td>
<td>STAY PLAY SWEET HAIR ENGAGING PRACTICING BEER CAMERA PASSION</td>
</tr>
<tr>
<td>UNDERSTAND MUSIC DESPERATELY</td>
<td>CLUMPY CAR LEGS ACRONYMS SAY</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>internet</td>
<td>D URL LOL ROFL O</td>
</tr>
<tr>
<td>Zooniverse</td>
<td>ZOO GZ ZOITES SERENGETI ZOOITE OBJID OOTD</td>
</tr>
<tr>
<td>time</td>
<td>NIGHT MOMENT FUTURE MORNING TIME EVENING TOMORROW YEARS SEASON DURATION TONIGHT FRIDAY PERIOD SATURDAY</td>
</tr>
<tr>
<td>animals</td>
<td>WILDEBEEST ZEBRA CAT HOUSE-SIZED CATS HYENA HORNS HARTBEEST GAZELLE ANIMALS BUFFALO BIRD BUCK WARTHOG LION ANIMAL ZEBRAS</td>
</tr>
<tr>
<td>other science</td>
<td>NUCLEUS ATOM UNITS CANDIDATES SPECTRAL INHABITS ELLIPTICALS TELESCOPE QUANTUM VOLCANOES EIGENSTATE ATOMS PHYSICS CA GAS EQUILIBRIUM CLASSIFIERS PARTICLES PHOTONS</td>
</tr>
<tr>
<td>interaction</td>
<td>CIAO HIYA OH GOODNIGHT BYE YEAH HEY NIGHTY HEH HA</td>
</tr>
<tr>
<td>numbers</td>
<td>BILLION SEVENTY MILLION HUNDRED HALF BILLIONS THOUSAND</td>
</tr>
<tr>
<td>quotes</td>
<td>TARE TUTTARE SOHA OM TURE CLARKE ARTHUR PÆANS AMARA MILLIWAYS STRIP-MINE SERAFINOWICZ KLEE BOWING KANO JIGARO EXALTED FORSTER HOWARDS</td>
</tr>
<tr>
<td>other</td>
<td>AM THEORY WELL COOKIES SPIDERS SPIDER ROSE SPIN MODEL INTERACTING EXPANSION TALK IRREGULARS RING C Z MAG SM EST FERMATS IV ES B Q AGIAN K MO ARM</td>
</tr>
</tbody>
</table>

Table 3: Selected KeyWords when 'topsix' is compared with 'bottomseven'

Table 3 shows a selection of the 500 KeyWords that were generated when 'topsix' was compared against 'bottomseven'. The KeyWords were manually categorised into the 14 categories shown. Space is one of the largest categories with 56 KeyWords.\(^1\)

This is clearly influenced by the subject matter of one of the largest and most successful Zooniverse projects - Galaxy Zoo. Though unsurprising as KeyWords, one should not hastily overlook such a category as uninteresting, however, because it highlights what subjects the volunteers are choosing to focus on, and could potentially be used to predict the success of a new project. Similarly the animals

\(^1\) Only a subset of the larger categories were reproduced in table three.
listed are influenced by the subject matter of a popular project - Snapshot Serengeti - but the names of animals and related words coming through as KeyWords could indicate animals that are relatively easy to identify and may provide Zooniverse leaders with useful information about how to adjust their training materials to support discussion and identification of less well-known animals.

Other sections such as grammar, internet and interaction are less obviously influenced by the subject matter and provide a snapshot of the linguistic behaviour of volunteers interacting and working together. The set of personal pronouns we, me, you, I, she (as well as Im and ya as a variant of you) suggest a friendly atmosphere where volunteers are engaging in personal chat as well as discussing science. Lines 1, 2 and 3 below show the use of ya in the three samples of text from top six.

(1) Well CIS Miami is on, so see ya all later. ;)
(2) a monocle, much cooler :D c ya Dave ;D ;D
(3) you a Horlicks :) That helps ya sleep. Well it is cloudy, but

Caution is needed when it comes to interpretation of any results based on this method. I am not claiming that a project leader can simply use such words more frequently to improve interaction and better statistics in terms of volunteers classifying their data. It is possible that other factors such as regular interaction between professional scientists and volunteers keeps volunteers interested, and that any linguistic evidence of a friendly atmosphere comes as a secondary effect. The linguistic atmosphere briefly sketched out here highlights areas of variation and themes that I will discuss further in the paper. Our discussion around this topic and further work could prove valuable to the leaders of other online communities who want to understand the relationship between project goals and online interaction between site users.

References


http://www.port.ac.uk/centre-for-european-and-international-studies-research/research-projects/language-of-citizen-science/
A corpus-based discourse analysis of the representation of trans people in the British press from 2013 to 2015
Angela Zottola (University of Napoli Federico II, Italy)

As demonstrated by numerous scholars (Baker and McEnery 2005; Bednarek 2006; Baker 2006, 2010, 2012; Baker, Gabrielatos & McEnery 2013; Partington 2015), Corpus Linguistics has proven to be a very useful and, at times, necessary tool for the analysis of large amounts of data in order to uncover linguistic and semantic patterns in the representations of social categories, minority groups or even individuals who managed to attract the attention of society upon them. Therefore, the purpose of this research is to highlight those linguistic and semantic patterns utilized in the representation of transgender people in the British press.

In the last decade, the construction of transgender people’s identity and the representation of their community has been increasingly raising interest in scholars in the field of (socio)linguistics (Kulick 1999; Zimman 2009, 2010, 2014; Baker 2014) and, more in general, in society as a whole, due to the considerable attention drawn by public episodes in the social, cultural and legal spheres, such as the transition of former athlete Caitlyn Jenner, in the U.S.A, the various trans soldiers involved in public affairs in the last few years, as the Chelsea Manning case, or even the complex issue regarding gender neutral toilets, which raised discussions globally.

In a society where gender diversity is a highly discussed topic, language, due to its social function, may take on a significant role in shaping and representing new gendered communities of practice. The existing binary and heteronormative linguistic categories, generally used in defining gender, are clashing with the current, so far unrepresented and now emerging communities, possibly leading to the creation of new hybrid, inclusive, non-discriminating discourses that comprise social, cultural and legal issues. On the basis of this popularity, the press works as one of the most active actors in the creation of these discourses; therefore, it became the primary focus for the collection of the corpus under scrutiny in this investigation. The geographical area investigated, the United Kingdom, was chosen as it represents one of the countries that first introduced the world to the discussion about transgender equality as a consequence of the abovementioned Manning case. With regards to these assumptions, this study has the scope of investigating how transgender individuals and their communities are shaped in and through language, starting from the premise that language may profoundly influence the way gender is understood by society, being one of the most powerful means of manipulation, through which it is possible to persuade and even instill specific beliefs or convictions.

The corpus collected so as to pursue the aim of this investigation comprises eight British national newspapers, namely The Guardian, the i, The Daily Telegraph, The Times, the Daily Express, the Daily Mail, the Daily Mirror and The Sun. The newspapers were chosen according to their circulation rates, retrieved from the UK Audit Bureau of Circulation (ABC 2015). Apart from distribution percentages, the other criterion followed in the selection of the newspapers was that of equally representing one of the dichotomies the British press is often defined by, that is, the distinction between quality and popular press (Jucker 1992). In fact, the first four newspapers can be considered as representative of the first category, while the others can be seen as belonging to the second. The news articles were collected in a time span that stretches from January 2013 to December 2015. Except for rubrics advertising the weekly television, cinema and theatre schedules,
different genres of news articles were included in the corpus (i.e., news stories, editorials, etc.), for a total of 3,119 news articles and over 2 million word tokens. For the purpose of this analysis, the TransCor (i.e., entire corpus) was divided into two subcorpora: the QualCor and the PopCor, each comprising all news articles from the quality press and the popular press respectively. Nonetheless, the analysis was conducted not only according to this distinction but diachronically as well.

Preliminary findings underline a difference in the representation depicted by the quality and the popular press, specifically in the use of terminology and, generally speaking, on the semantic prosody surrounding the topic. In fact, the collocation analysis highlights, for example, that while in many cases trans people are depicted as victims in in the QualCor, they become the perpetrators of violence in the PopCor.

Moreover, despite the very short time span covered by the TransCor, one of the most impressive results, emerging from the analysis of the two subcorpora, was represented by the changing in use surrounding the adjectives ‘transgender’ and ‘transsexual’, shifting towards a more inclusive and unbiased rhetoric in the last year (2015). The investigation has underlined that, the two terms ‘transgender’ and ‘transsexual’ were initially used both as adjectives and as nouns, a function, the latter, which is disliked by the majority of the people belonging to the trans community. This trend tends to decrease in the following years, and the use of the two adjectives, erroneously used as synonyms and as nouns, starts to be used correctly more and more often.

Keywords were also an object of investigation. In order to compute the keyword lists for the QualCor and the PopCor, the BE06 corpus (Baker 2009) was used as reference corpora. This technique brought forward some other topics associated with trans people that did not emerge in the collocation analysis, such as the term ‘sex’.

The research has used as its main corpus analysis tool AntConc (Anthony 2014), which has allowed to both perform the keyword analysis previously introduced and further explore the collocational patterns associated with the word displaying the highest degree of keyness.

In the framework of Critical Discourse Analysis (Fairclough 1995, 2005; Wodak 1997; Bucholtz 2003; van Leeuwen 2005) and through the use of Corpus Linguistics methodologies (McEnery and Wilson 1996; Baker 2006, 2014; McEnery and Hardie 2012), this study will present the final outcomes of a research that focuses on the representation of transgender people in the British press from the point of view of their depiction as social actors, the strategies of legitimization employed in this representation (van Leeuwen, 2007) and the semantic prosody (Sinclair 1991; Louw 1993; Stubbs 1996; Partington 1998; Hunston 2007; Partington, Duguid and Taylor 2013) constructed discursively in the description of trans social actors.

References


Uncovering Informal Expressions Used in English NNS Thesis Writing
Hao-Jan Howard Chen (National Taiwan Normal University, Taiwan)

Hyland and Jiang (2016) explored the perception that since informality has now invaded a large range of written and spoken domains of discourse, academic writing has also followed this trend to accept more informal expressions. They, however, found that only a small increase in the use of some informal features and academics in general still hold high standards toward formality in academic writing.

Given that the informal expressions are widely used in a wide range of written and spoken discourses, many L2 learners were found to use various informal expressions in formal academic writing, as pointed out by Gilquin and Paquot (2008). However, few empirical studies systematically investigated these great challenges faced by L2 writers. To better understand L2 learners’ real difficulties in using these various informal features, this study compared two NS and NNS academic corpora to uncover informal expressions used by NNS writers. The NNS academic corpus is consisted of 210 Taiwanese graduate students’ theses in computer science. The NS corpus includes 171 papers from well-known international journals of computer science. By comparing these two sets of data, it is possible to pinpoint the differences between novice academic writers and professional writers in computer science.

There are various possible methods we can further compare the NNS and professional academic corpora. In this study, the research focus would be on the unusual high-frequency informal expressions used by NNS academic writers. It is expected that some of these high-frequency informal expressions can be identified from the learner corpus. Keyword and key clusters identification method was found useful by several researchers (Mahlberg, 2007; Fischer-Starcke, 2009). In this study, WordSmith 7 created by Mike Scott was adopted and the keyword and key clusters extraction functions were used. The word list and the clusters (2-word, 3-word, and 4-word clusters) of these two corpora were first compiled. Then these lists produced by professional and NNS academic writers were compared using the keyword tool of WordSmith.

First, by running the keyword analysis, several informal items were uncovered from the NNS academic corpus. These include connector ‘besides’ and several informal verbs such as ‘get’, ‘wants’, ‘means’. Some more examples are shown in Table 1.

<table>
<thead>
<tr>
<th>Key word</th>
<th>Freq. (Professional)</th>
<th>RC. Freq. (Learners’ Theses)</th>
<th>Log_L</th>
</tr>
</thead>
<tbody>
<tr>
<td>besides</td>
<td>37</td>
<td>344</td>
<td>-242.99</td>
</tr>
<tr>
<td>according</td>
<td>512</td>
<td>1188</td>
<td>-187.62</td>
</tr>
<tr>
<td>get</td>
<td>262</td>
<td>726</td>
<td>-164.87</td>
</tr>
<tr>
<td>wants</td>
<td>23</td>
<td>206</td>
<td>-142.90</td>
</tr>
<tr>
<td>lot</td>
<td>32</td>
<td>189</td>
<td>-102.05</td>
</tr>
<tr>
<td>means</td>
<td>440</td>
<td>838</td>
<td>-75.22</td>
</tr>
</tbody>
</table>

Table 1. Some Informal Words Used by NNS Academic Writers

For the 2-word clusters, some informal phrasal verbs such as ‘find out’ were found. Some informal verbs such as ‘wants’ and ‘get’ were also widely used in this learner corpus. These items were in general informal but many Taiwanese academic writers were not aware of their informality. In addition, in the theses written by graduate students, there are many
examples of first personal plural pronoun ‘we’. These include ‘we can’, ‘our system’, ‘we proposed’, ‘we will’. Table 2 shows some of the key differences between NS and NNS.

<table>
<thead>
<tr>
<th>Key word</th>
<th>Freq. (Professional)</th>
<th>RC. Freq. (Learners' Theses)</th>
<th>Log_L</th>
</tr>
</thead>
<tbody>
<tr>
<td>we can</td>
<td>678</td>
<td>1730</td>
<td>-336.09</td>
</tr>
<tr>
<td>our system</td>
<td>19</td>
<td>318</td>
<td>-279.58</td>
</tr>
<tr>
<td>wants to</td>
<td>15</td>
<td>200</td>
<td>-163.65</td>
</tr>
<tr>
<td>get the</td>
<td>51</td>
<td>271</td>
<td>-134.88</td>
</tr>
<tr>
<td>find out</td>
<td>10</td>
<td>153</td>
<td>-130.97</td>
</tr>
<tr>
<td>lot of</td>
<td>22</td>
<td>173</td>
<td>-112.11</td>
</tr>
<tr>
<td>we proposed</td>
<td>19</td>
<td>161</td>
<td>-108.60</td>
</tr>
<tr>
<td>it means</td>
<td>11</td>
<td>131</td>
<td>-102.87</td>
</tr>
<tr>
<td>a lot</td>
<td>32</td>
<td>184</td>
<td>-97.40</td>
</tr>
<tr>
<td>we will</td>
<td>429</td>
<td>872</td>
<td>-96.71</td>
</tr>
<tr>
<td>can get</td>
<td>12</td>
<td>123</td>
<td>-90.83</td>
</tr>
</tbody>
</table>

Table 2. Some 2-word Informal Expressions Used by NNS Academic Writers

For 3-word clusters, some informal expressions such as ‘a lot of’, ‘reason is that’, ‘and so on’, and ‘deal with the’ were found. More examples can be found in Table 3. Similar to what we found in the 2-word clusters, there are many expressions containing first personal plural pronoun ‘we’. It seems that graduate students were giving credits to the whole research team/lab and many also believe that the thesis work should be shared with their thesis advisers.

<table>
<thead>
<tr>
<th>Key Clusters</th>
<th>Freq. (Professional)</th>
<th>RC. Freq. (Learners' Theses)</th>
<th>Log_L</th>
</tr>
</thead>
<tbody>
<tr>
<td>a lot of</td>
<td>22</td>
<td>169</td>
<td>-108.09</td>
</tr>
<tr>
<td>we have to</td>
<td>25</td>
<td>148</td>
<td>-80.05</td>
</tr>
<tr>
<td>we can see</td>
<td>57</td>
<td>194</td>
<td>-60.54</td>
</tr>
<tr>
<td>reason is that</td>
<td>14</td>
<td>97</td>
<td>-58.34</td>
</tr>
<tr>
<td>we can find</td>
<td>13</td>
<td>93</td>
<td>-57.05</td>
</tr>
<tr>
<td>to get the</td>
<td>20</td>
<td>99</td>
<td>-46.41</td>
</tr>
<tr>
<td>we propose a</td>
<td>74</td>
<td>199</td>
<td>-42.79</td>
</tr>
<tr>
<td>and so on</td>
<td>55</td>
<td>152</td>
<td>-34.35</td>
</tr>
<tr>
<td>we can use</td>
<td>32</td>
<td>107</td>
<td>-32.60</td>
</tr>
<tr>
<td>we will discuss</td>
<td>11</td>
<td>62</td>
<td>-32.33</td>
</tr>
<tr>
<td>we set the</td>
<td>15</td>
<td>70</td>
<td>-31.12</td>
</tr>
<tr>
<td>deal with the</td>
<td>21</td>
<td>82</td>
<td>-30.33</td>
</tr>
<tr>
<td>we can observe</td>
<td>15</td>
<td>66</td>
<td>-27.72</td>
</tr>
</tbody>
</table>

Table 3. Some 3-word Informal Expressions Used by NNS Academic Writers

For 4-word clusters, some informal expressions such as ‘to find out the’, ‘it means that we’, ‘reason is that we’, ‘that is to say’, and ‘a lot of time’ were found. More examples can be found in Table 4. Similar to what we found in the other groups of word clusters, there are many expressions containing first personal plural pronoun ‘we’. As pointed out in the previous section, students chose to give credits to the whole research team/lab or their thesis advisers. Thus, many chose to use ‘we’ in their thesis writing.
Based on the preliminary findings, it is clear that many L2 novice writers were not fully aware of register differences, and they tended to use some informal expressions in their formal thesis writing. It is thus important to prepare some learning materials for these EAP writers. Academic writing instructors can then use these materials and exercises targeting on informality issues to draw L2 learners’ attention to these informal items.

**References**


Corpora, Prototypes and Literary Translation
Natalie Finlayson (University of Glasgow, UK)

Project Summary

How can corpora be used to enhance the processes by which we determine which lexical items are prototypical in a language, and to what extent do prototypes vary interlingually? Are such differences reflected in translated literature, and could a knowledge of internal category structure be of benefit to literary translators? My Ph.D. research suggests that these questions may be fruitfully addressed from a corpus linguistic perspective, combining monolingual corpus-based methods with informant-based research to identify prototypes in British English and Standard German, and using these as a lens through which to analyse translation strategies.

Corpora and Prototypes

The poster shows how traditional methods of measuring internal category structure (see Battig & Montague, 1969; Hampton & Gardiner, 1983; Glauer et al., 2007) may be modernised using corpus methodologies. It argues that the content of the British National Corpus and *Digitales Wörterbuch der deutschen Sprache* reveal far more about the frequency of use of lexical items in everyday language and the ideological attitudes and concerns (Anderson & Corbett, 2017, p. 34) of speakers than a study of probability of production could. It demonstrates that corpus-based results may be combined with those of an informant-based study (modelled on Rosch’s (1973, 1975) work on goodness-of-example) to create an average measure of prototypicality, contributing to a cognitive framework from within which to analyse translation strategies.

Corpora and Literary Translation

By drawing parallels between the role of literary translators as ‘couriers of culture’ (Landers, 2001, p. 72), and the corpus linguistic view of language as reflecting and reproducing culture across generations (Laviosa, 2002, p. 8), it is argued that these complementary disciplines may be combined to examine current language in literary texts and their translations from the perspective of the experiences, cultures and worldviews of individual language users.

The poster depicts the role of *Tetrapla* (Woolls, 2008-2015), a software program optimised for the analysis of English-German-English translations, in identifying translation strategies in a custom-made parallel corpus of popular fiction. By referring to prototypicality values assigned to individual lexical items, translation phenomena are discussed from the perspective of prototype theory, which could explain why a *moulting* old feather duster, for example, may be described using prototypical equivalent *zerzaust* (unkempt), rather than lexical equivalent *mausernd* (moultling), in translation. This example gives weight to Aitchison’s suggestion that it may be more important to replace a prototypical instance with another prototypical
instance rather than with an exact lexical equivalent’ (1996, p. 18), and the extent to which this is the case more generally is discussed with reference to further instances.

**Implications**

Results have the potential to make a contribution to the underexplored area of prototypicality in applied linguistics (Geeraerts, 2010, p. 255), and to facilitate broader discussions of interlingual differences in semantic structure. In practice, findings may equip literary translators with an awareness of prototypicality, which could contribute something of value in terms of cultural knowledge exchange at a time when interest in translated fiction in the UK is at an all-time high (Flood, 2016).

**Reference List**


Concessive conditional sentence differs from ordinary conditional sentence in the sense that the main clause of the former takes place whether the subordinate clauses is fulfilled or not, while the main clause of ordinary conditional sentence will only take place if the subordinate clause is fulfilled. (König, 1986; Buckley, 2004). In Arabic, there are five common compound particles that express the concessive conditional meaning; they are: wa-in, wa-law, ḥattā wa-in, ḥattā law and ḥattā wa-law (Badawi et al. 2004). In modern Arabic grammar books, it has been claimed that the propositions that follow these particles can be rendered into two possible modality meanings: factual and non-factual. In the former, the particle can be equivalent to the English concessive conjunction ‘even though’, while in the latter, the particle can be equivalent to the English concessive conditional conjunction ‘even if’ (Cantarino, 1975, vol.3; Badawi et al. 2004; Buckley, 2004). However, to my knowledge, no prior study has been undertaken to empirically investigate the frequency of these particles in Arabic written texts and their interaction with the two modality meanings mentioned above.

In this paper, I will employ a corpus-based methodology to explore how the concessive conditional particles are actually used in Modern Literary Arabic (MLA) discourse to convey factual and hypothetical (i.e. non-factual) meanings. This will mainly focus on observing the frequency of each particle and the frequency of their meanings. The use of the corpus-based method and technique in this study is justified by the fact that they “allow us to determine common and uncommon choices and to see the patterns that reveal what is typical and untypical in particular contexts” (Conrad, 2012: 228).

The data sample is drawn from the Brigham University electronic Arabic Corpus tool developed by Dilworth Parkinson (http://arabicorpus.byu.edu/). The total number of the words in the whole corpus is 173,600,000. The Modern Literature section, which is the target discourse, is composed of 32 texts which jointly contain 1,026,171 words. Using the basic search query and the concordance lines, the total number of concessive conditional particle tokens obtained is 340. As a first step, the frequencies of the particles have been determined. The following table shows the results of this step:

<table>
<thead>
<tr>
<th>Particle</th>
<th>Frequency</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>wa-in</td>
<td>92</td>
<td>27 %</td>
</tr>
<tr>
<td>wa-law</td>
<td>149</td>
<td>44 %</td>
</tr>
<tr>
<td>ḥattā wa-in</td>
<td>13</td>
<td>4 %</td>
</tr>
<tr>
<td>ḥattā law</td>
<td>65</td>
<td>19 %</td>
</tr>
<tr>
<td>ḥattā wa-law</td>
<td>21</td>
<td>6 %</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>340</strong></td>
<td><strong>100 %</strong></td>
</tr>
</tbody>
</table>
As shown in the table, the particle *wa-law* is the dominant one in the data, followed by the particle *wa-in* in second place. As a second step, a qualitative analysis of the examples will be incorporated in order to contextually identify the Modality meanings of each particle, accompanied with statistical descriptions.

**References**


A Corpus Approach to the Disciplinary History of Second Language Writing
Zhaozhe Wang and Ge Lan (Purdue University, US)

The field of second language (L2) writing has grown into a legitimate field of disciplinary inquiry within a relatively short period of time. Yet despite disciplinary historians’ endeavor to provide a narrative of the evolution and shape the identity of the field of second language writing (Matsuda, 1998, 1999, 2006, Silva, 1990, 1993, 2004, etc.), historical inquiry is still largely overshadowed by empirical inquiry, as suggested by the latter’s dominant presence in the flagship journals in the field (Canagarajah, 2016). Nevertheless, historical inquiry plays a pivotal role in shaping and contextualizing our understanding of the field. The integration of computer technology into corpus linguistics makes feasible the exploitation of massive corpora (McEnery, Xiao, and Tono, 2006). New research methods such as corpus analysis with the aid of digital technology enable researchers to attain a higher level of historical detachment, as large collections of texts can be machine analyzed and interpreted for consistency and reliability (Biber, Conrad, and Reppen, 1998). Therefore, to foreground disciplinary historical study in the digital era, drawing on meta-analysis and corpus linguistics methodologies, we introduce and demonstrate a new approach—a corpus-driven historical study, to examine and complement the existing historical narratives of the field of second language writing. The research questions we pose are (1) How, if possible, can the use of corpus analysis illuminate our understanding of the historical construction of the field of Second Language Writing? (2) Does the change in frequency of certain high-frequency collocates reflect the shift of scholarly attention in the field across different periods of time? If so, how?
<table>
<thead>
<tr>
<th>Theme</th>
<th>Collocates</th>
<th>Freq_1980s</th>
<th>Freq_1990s</th>
<th>Freq_2000s</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Process</strong></td>
<td>Peer review</td>
<td>6.32</td>
<td>189.51</td>
<td>143.04</td>
</tr>
<tr>
<td></td>
<td>Writing task</td>
<td>85.35</td>
<td>98.41</td>
<td>262.34</td>
</tr>
<tr>
<td></td>
<td>Writing process(es)</td>
<td>222.45</td>
<td>376.57</td>
<td>359.12</td>
</tr>
<tr>
<td></td>
<td>Composing process(es)</td>
<td>417.27</td>
<td>159.41</td>
<td>93.84</td>
</tr>
<tr>
<td><strong>Text</strong></td>
<td>Academic writing</td>
<td>360.36</td>
<td>412.36</td>
<td>222.78</td>
</tr>
<tr>
<td></td>
<td>Contrastive rhetoric</td>
<td>69.55</td>
<td>346.48</td>
<td>251.99</td>
</tr>
<tr>
<td></td>
<td>Error correction</td>
<td>41.10</td>
<td>83.77</td>
<td>273.90</td>
</tr>
<tr>
<td></td>
<td>Academic discourse</td>
<td>72.71</td>
<td>176.49</td>
<td>88.87</td>
</tr>
<tr>
<td></td>
<td>Discourse analysis</td>
<td>18.97</td>
<td>39.85</td>
<td>118.08</td>
</tr>
<tr>
<td><strong>Literacy</strong></td>
<td>Writing proficiency</td>
<td>208.64</td>
<td>102.48</td>
<td>78.52</td>
</tr>
<tr>
<td></td>
<td>Writing ability</td>
<td>56.90</td>
<td>96.79</td>
<td>72.43</td>
</tr>
<tr>
<td></td>
<td>Academic literacy</td>
<td>0.00</td>
<td>62.63</td>
<td>75.48</td>
</tr>
<tr>
<td></td>
<td>Critical thinking</td>
<td>15.81</td>
<td>117.12</td>
<td>28.00</td>
</tr>
<tr>
<td><strong>Language</strong></td>
<td>Linguistic features</td>
<td>0.00</td>
<td>12.20</td>
<td>121.74</td>
</tr>
<tr>
<td></td>
<td>Native language</td>
<td>154.90</td>
<td>95.16</td>
<td>110.78</td>
</tr>
<tr>
<td></td>
<td>Language proficiency</td>
<td>98.00</td>
<td>143.96</td>
<td>81.56</td>
</tr>
<tr>
<td><strong>Context</strong></td>
<td>Discourse community</td>
<td>12.64</td>
<td>154.53</td>
<td>183.22</td>
</tr>
<tr>
<td></td>
<td>Cross-cultural</td>
<td>34.77</td>
<td>145.59</td>
<td>83.89</td>
</tr>
<tr>
<td><strong>Writer</strong></td>
<td>Native speakers</td>
<td>344.57</td>
<td>243.19</td>
<td>153.39</td>
</tr>
<tr>
<td></td>
<td>L2 writers</td>
<td>15.81</td>
<td>238.31</td>
<td>331.73</td>
</tr>
<tr>
<td></td>
<td>ESL writers</td>
<td>123.29</td>
<td>311.51</td>
<td>54.78</td>
</tr>
<tr>
<td></td>
<td>International students</td>
<td>50.58</td>
<td>109.80</td>
<td>49.30</td>
</tr>
<tr>
<td><strong>Feedback</strong></td>
<td>Teacher feedback</td>
<td>0.00</td>
<td>181.37</td>
<td>175.91</td>
</tr>
<tr>
<td></td>
<td>Peer feedback</td>
<td>6.32</td>
<td>150.47</td>
<td>123.56</td>
</tr>
<tr>
<td></td>
<td>Corrective feedback</td>
<td>15.81</td>
<td>12.20</td>
<td>157.04</td>
</tr>
<tr>
<td></td>
<td>Peer response</td>
<td>69.55</td>
<td>303.37</td>
<td>156.43</td>
</tr>
<tr>
<td></td>
<td>Error feedback</td>
<td>22.13</td>
<td>3.25</td>
<td>169.82</td>
</tr>
</tbody>
</table>

Table 1. Normalized frequencies of the theme-based collocates

Specifically, we quantitatively and qualitatively investigate a collection of high-frequency keywords and collocates representing frequently referred terminologies in a corpus consisting of 415 journal articles (total token: 3,186,228) that address second language writing issues in 4 leading journals in the interdisciplinary field, including Journal of Second Language Writing, TESOL Quarterly, College Composition and Communication, and Written Communication. The frequencies of the keywords and collocates in the corpus are compared across 3 decades (1980s, 1990s, and 2000s) to track the shift of scholarly attention over time. The “shift of scholarly attention” is operationalized as the change of frequency of the high-frequency collocates based on the high-frequency high-keyness keywords in the scholarship during a certain decade. We present, analyze, and contextualize the frequency change of these collocates categorized into 7 thematic groups—“process,” “text,” “literacy,” “language,” “context,” “writer,” and “feedback.” The findings suggest that the terminological trends in the scholarship are generally aligned with the historical narrative, and a plethora of emerging terminologies are replacing the ones that used to display a dominant presence in the literature.
References


Introduction

This paper aims to explore the issue of competitiveness in the ‘graduate job market’ and the notion of ‘employability’, as presented by Universities in the UK through their career services webpages. It is part of a wider project that seeks to understand the role of careers services inside academia. With a combination of Critical Discourse Analysis methods and Corpus Linguistics tools, the analysis focuses on identifying, interpreting, explaining and evaluating (Baker and McEnery, 2015, pp. 2-3) the ‘reality’ presented by these services as students are expected to ‘invest’ their time at university in getting prepared for the transition from HE to the workplace.

HE in the UK and the concept of ‘employability’

Since the early 1980s, UK governments have introduced numerous reforms to the educational system. By the mid-1990s, ‘universities were considered by the policy makers to be more about conferring private benefit upon individuals than public benefit upon society as a whole’ (Williams, 2013, p. 41). Emphasis on ‘the individual’ is also clearly expressed in the Browne review:

The primary beneficiary of higher education is the individual student. The student chooses where to study and what to study; and the student chooses where to use the new skills they have acquired. Businesses benefit from employing highly skilled graduates and they pay for that benefit through higher wages (2010, p. 54).

Besides the fact that the individual is considered the primary beneficiary of higher education and the one who has the ‘power’ to choose (the power of the consumer), this statement creates a direct link between higher education, students, and employers which all have one thing in common: the interest in teaching, acquiring and using (respectively) ‘skills’.

The concept of ‘employability’ promoted by powerful groups, such as governments, business organisations, employers, HE institutions, and the media, emphasises the development of employability skills as necessary for those who wish to enter the ‘global graduate job market’. The discourse of employability promotes a reality where economic insecurity and labour competitiveness are presented as natural, as ‘common sense’ (Fairclough, 2015). Individuals are considered responsible for their own success in the labour market and governments become the ‘enablers’ that can provide the possibility to ‘make necessary choices to become employable’ (Fejes, 2010, p. 99). As Boden and Nedeva (2010, pp. 43-44) note, the UK government has ‘appropriated’ itself from any blame about the ‘market dysfunction’ while universities were given the responsibility of preparing the ‘workers’ employers need. In other words, universities have taken up the task of preparing students to enter the ‘job market’.
More specifically, *Careers and Employability* services, which are considered a specialized branch inside the HE institutional system, aim to guide and help students towards becoming employable during their time at university. The concept of ‘employability’ in UK universities suggests that if HE students focus on ‘gaining’, ‘developing’, ‘enhancing’, ‘improving’ their skills while at university, they will become more employable and thus earn an advantage in ‘securing’ a job after graduation. It is thus perceived as a remedy to the problem of unemployment, especially in times of economic ‘instability’, where individuals are expected to adapt and become flexible to the needs of the ‘global market’ should they wish to become employed.

**Methodology and Corpus**

In line with previous research that follows a critical approach on the marketization of HE, such as Fairclough (1993); Mautner (2005, 2010); Askhave (2007); Mayr (2008); Zhang and O’Halloran (2013), this paper aims to critically examine the discourse used by careers services in UK universities. More specifically, it focuses on ‘competition’ and the notion of ‘employability’. The study of language and its use in contemporary capitalism can unveil the promotion and circulation of ideologies that intend to influence people’s perception of how the world works and also affect their choices and actions.

This corpus-based critical discourse analysis aims to identify and interpret linguistic patterns of the search terms ‘competition’, ‘competitive’, ‘employability’, and ‘employable’, using AntConc’s *concordance* and *collocates* tools (Anthony, 2014). The *Careers and Employability Webpages Corpus* (CEW15) was built in 2015 and consists of 2.6 million words collected from 58 UK university websites and in particular their *Careers Services* sections.

**Analysis**

‘**Stand out from the crowd**’ in the ‘**competitive graduate job market**’

Using AntConc’s KWIC function, the analysis begins by looking closely at the terms ‘competition’ and ‘competitive’. The idea of ‘competition’ in the context of employment, or the ‘job market’ as expressed by the careers services, is described as ‘fierce’, ‘high’, ‘intense’, ‘stiff’, ‘strong’, and ‘tough’ (Figure 1).

![Figure 1: Concordance Lines - 'competition' in CEW15](image-url)
Similarly, the adjective ‘competitive’ is used to characterise the availability of employment positions. Students and graduates are informed of the ‘competitive employment market’, ‘competitive global economy’, ‘competitive graduate job/labour market’. It could be argued that the rhetoric of ‘fierce’ competitiveness addressed to students or graduates who are searching for employment, can act as stimulus to take ‘appropriate’ action. Thus, students are urged to gain a ‘competitive advantage’ or a ‘competitive edge’ in the ‘graduate job market’. They are also encouraged to ‘stand out from the competition’ or ‘the crowd’. The close examination of the pattern ‘stand out from’ - found 143 times in the corpus - shows that students are expected to ‘improve’, ‘enhance’, and ‘boost’ their CVs and skills, ‘take action early’, get ‘work experience’ and ‘invest’ in their ‘employability’ if they want to ‘secure’ a place in the ‘competitive job market’.

The discourse of employability

The analysis then moves on to examine the concept of ‘employability’. AntConc shows that within a span of -3 +3 words, the most frequent lexical collocates of the search term ‘employability’ are: ‘careers’, ‘skills’, ‘service’, and ‘team’ (Table 1):

<table>
<thead>
<tr>
<th>Freq</th>
<th>Freq(L)</th>
<th>Freq(R)</th>
<th>MI</th>
<th>Collocate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1593</td>
<td>863</td>
<td>730</td>
<td>3.48395</td>
</tr>
<tr>
<td>2</td>
<td>1504</td>
<td>1146</td>
<td>358</td>
<td>3.27564</td>
</tr>
<tr>
<td>3</td>
<td>984</td>
<td>916</td>
<td>68</td>
<td>5.53015</td>
</tr>
<tr>
<td>4</td>
<td>683</td>
<td>473</td>
<td>210</td>
<td>2.31987</td>
</tr>
<tr>
<td>5</td>
<td>585</td>
<td>41</td>
<td>544</td>
<td>5.25245</td>
</tr>
<tr>
<td>6</td>
<td>578</td>
<td>467</td>
<td>111</td>
<td>3.56048</td>
</tr>
<tr>
<td>7</td>
<td>426</td>
<td>290</td>
<td>136</td>
<td>2.25812</td>
</tr>
<tr>
<td>8</td>
<td>391</td>
<td>37</td>
<td>354</td>
<td>5.60622</td>
</tr>
<tr>
<td>9</td>
<td>349</td>
<td>34</td>
<td>315</td>
<td>6.19010</td>
</tr>
<tr>
<td>10</td>
<td>313</td>
<td>133</td>
<td>180</td>
<td>2.21464</td>
</tr>
</tbody>
</table>

Table 1: 'employability' collocates

In recent years, careers services in UK universities have integrated the term ‘employability’ in their services’ titles (‘Careers and Employability Service’), professional titles (‘Careers and Employability Adviser’, ‘Careers and Employability team’, ‘the Global Employability Team’), events (‘Careers & Employability Fair’), and resources used (Careers & Employability booklet’). In addition, students are instructed to ‘develop’, ‘enhance’, ‘identify’ or even ‘sell’ to employers a plethora of ‘employability skills’ while studying at university.

As shown in Table 1, one of the most frequent collocates of ‘employability’ that can provide information about the ideological stances that follow this term, is the possessive determiner ‘your’. A close analysis of the concordance lines (‘your employability’) shows that students are directed to ‘boost’, ‘build’, ‘demonstrate’, ‘enhance’, ‘improve’, ‘increase’, ‘shape’ and ‘support’ their employability.

Similarly, the examination of the adjective ‘employable’, found 157 times in the corpus, shows that the services highlight their own helping role in making students become ‘more/highly/globally employable graduates’. Most importantly, however, the services suggest that the students should take action and make themselves employable (Figure 2):
Discussion and conclusions

The discourse of careers services in UK universities is in line with the neoliberal thought and rhetoric that chooses to focus on an insecure and competitive way of living without really explaining the reasons that have led society and its economy in this problematic state. The concept of ‘employability’ is promoted as a solution to this challenging reality that students/graduates are asked to follow blindly should they wish to get employed. Students are urged to ‘stand out from the crowd’ and make themselves employable. Is this, however, the solution that will actually help young people in the long term? Who benefits from the concept of ‘employability’? Is becoming ‘employable’ in order to compete in this ‘tough’ and ‘fierce’ ‘graduate job market’, a fruitful solution or a pseudo-therapy to the problematic neoliberal reality that students are asked to deal with? And finally, do careers services ‘teach’ students how to ‘stand out from the crowd’ or become a part of the ‘job-hunting’ crowd? These are the ‘evaluating’ questions that the final part of the analysis wishes to raise which is the element that makes the discourse analysis critical (Baker and McEnery, 2015: p. 3).

References


A dictionary, a survey and a corpus walked into a courtroom...: An evaluation of resources for adjudicating meaning in trademark disputes
Olumide Popoola (Aston University, UK)

Introduction

When trademark disputes turn on issues of meaning and language use, courts have resorted to three linguistic tools. In the UK, dictionaries, judges' intuitions and market research are the key resources used; in the US expert linguists have developed a “trademark linguistics” (e.g. Butters, 2008; Shuy, 2002), utilising lexicographic methods on legal databases (e.g. LexisNexis, Westlaw) and, more recently, corpora such as COCA.

Corpus analytical tools have been established (e.g. Microsoft Corporation v. Apple Inc., 2010; Kilgarriff, 2015) to assess whether disputed trademarks such as 'Botox', 'Velcro' or 'App Store' are or have become generic everyday expressions. However, research on distinguishing imaginative from arbitrary usage of lexical words in trademarks has been restricted to the psycholinguistic domain (e.g. Hotta & Fujita, 2012). This paper demonstrates a syntactic-pragmatic corpus analytical method for determining the potential meaning of lexical brand names in trademark dilution disputes. The analysis presented here is compared to judges' lexicographical analysis and market research presented as evidence in court during a three-year dispute over the CARBON VIRGIN trademark (Michael Casey v. Virgin Enterprises Ltd, 2010/2011/2012). Corpus linguistics emerges as a viable alternative to both dictionaries and surveys.

Corpus Data

Lower case ‘virgin’ was analysed across three genre-specific corpora: The Birmingham Blog Corpus, Leeds University Corpus of British Newspapers, and a corpus compiled from Twitter searches for ‘virgin’ conducted over a two-week period.

<table>
<thead>
<tr>
<th>Table 1. Corpus data</th>
<th>Total number of words</th>
<th>Number of ‘virgin’ tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Birmingham Blog Corpus</td>
<td>600million</td>
<td>1721</td>
</tr>
<tr>
<td>Leeds University Corpus of British Newspapers</td>
<td>200million</td>
<td>865</td>
</tr>
<tr>
<td>Twitter</td>
<td>50million</td>
<td>1217</td>
</tr>
</tbody>
</table>
Analysis Methods

100 KWIC concordance examples of 'virgin' were randomly sampled from each corpus. These were first grouped by syntactic frames (see Table 2 below) and a register analysis was then conducted on examples within each frame utilising a Hallidayan framework with the parameters listed in Table 3 below.

Table 2: Syntactic frames for 'virgin' found in corpus.

<table>
<thead>
<tr>
<th>Frame Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIRGIN + Noun/Noun Phrase[NP]</td>
<td>&quot;virgin territory&quot;</td>
</tr>
<tr>
<td>Determiner/No Article/Quantifier + VIRGIN</td>
<td>&quot;the virgin&quot;</td>
</tr>
<tr>
<td>Noun/Noun Phrase[NP] + VIRGIN</td>
<td>&quot;kids' party virgin&quot;</td>
</tr>
<tr>
<td>Adjective + VIRGIN</td>
<td>&quot;famous virgin&quot;</td>
</tr>
</tbody>
</table>

Table 3: Register analysis framework

<table>
<thead>
<tr>
<th>Field</th>
<th>Topic</th>
<th>Connotation (Allan, 2007) summarized as positive/neutral/negative evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tenor</td>
<td>Communicative function based on Searle (1975) illocutionary act classification/</td>
<td></td>
</tr>
<tr>
<td>Mode (Halliday 1976, 2014)</td>
<td>Coherence; Text organisation (Given-New structure); Intertextuality (e.g. legal definition, pop song).</td>
<td></td>
</tr>
</tbody>
</table>

For comparison, both CARBON VIRGIN and a separately disputed trademark, VIRGIN THREADS, were analysed for plausibility within the NOUN/NP + VIRGIN and VIRGIN + NOUN/NP syntactic frames. Perceived plausibility levels were derived from the extent of fit between the disputed 'virgin' expressions and the uses of 'virgin' observed in the three corpora. Field, tenor and mode parameters were combined to generate a basic overall plausibility rating.

RESULTS

Figure 1: Plausibility ratings for disputed trademarks and suggested position on 'trademark distinctiveness spectrum' (Abercrombie & Fitch Co. v. Hunting World (1976))

<table>
<thead>
<tr>
<th>Disputed trademark</th>
<th>Field Plausibility</th>
<th>Tenor Plausibility</th>
<th>Mode Plausibility</th>
<th>Overall Plausibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>'CARBON VIRGIN'</td>
<td>Low/Null</td>
<td>Low/Null</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>'VIRGIN'</td>
<td>Low/Null</td>
<td>Medium</td>
<td>Low/Medium</td>
<td>Medium</td>
</tr>
</tbody>
</table>
Field

NOUN/NP + VIRGIN

'Virgin' as a noun postmodifier has the meaning of “naive in relation to” something other than sex. (The noun phrase 'sex virgin' is tautologous.) The examples of NOUN/NP + VIRGIN in the overall corpus all broadly relate to activities, pastimes and hobbies that one can do, know or experience e.g. attend an event (“Nationals virgin”), genre conventions (“hip-hop virgin”), enjoy a meal (“Nandos virgin”).

VIRGIN + NOUN/NP

Fixed collocations "virgin birth" and "virgin territory" aside, nominal postmodifiers of virgin generally fall into six categories: olive oil variants ("extra virgin olive oil"); commercial hair variants ("virgin hair"; "virgin weave"); nature features ("virgin land"; "virgin snow"); non-alcoholic cocktails ("virgin martini"); inexperienced or newbie ("virgin millionaires", "virgin blogger", "virgin (business) angel"); a state of innocence i.e. without prior exposure to unsocial acts (e.g. "virgin ears" for swearing)

Plausibility Evaluation

Since it is difficult to formulate a meaning of 'carbon' that can correspond to something that one can do, know or experience, the 'carbon virgin' rating for Field Plausibility is Low/Null. Although it is true that 'threads' is a synecdoche for clothes, none of the unclassified virgin word combinations (metaphors such as "virgin atmosphere") indicate a readily comprehensible meaning for "virgin threads". Consequently its Field Plausibility is Low/Null.

Tenor

NOUN/NP + VIRGIN

Instances of 'virgin' fall into two main interpersonal categories: i) representative self-descriptions (“I was a kids party virgin”) that do not require expert knowledge, and ii) directive communications delivered with the aim to persuade or advise, often used with positive connotation in PR/advertising contexts as a selling point (“worth dipping into for the mood-lighting virgin”). The use of hyphenated noun pre-modifiers (e.g. “buy-toilet virgin”, “FTSE-100-company-boss virgin”) is further evidence of sophisticated and witty coinage.

VIRGIN + NOUN/NP

Within a general environmental context, 'virgin' has both positive and negative connotations ("lure wealthy skiers to spend the day visiting virgin snow"); “The reason
we rape so many virgin sites with crude housing...”). It is also used persuasively e.g. in relation to holidays (“virgin sand”). Non-sexual references to personal physical features are also common as both self-representation (“My virgin eyes are gone now”), advice (“Those with weak stomachs or virgin ears should also stay away”) and negative evaluation (“virgin realms of incompetence”).

**Plausibility Evaluation**

CARBON VIRGIN’s scope for use as a personal description or in a persuasive context is limited by its specialised claimed meaning and lack of wit. Tenor Plausibility is Low/Null. Outside of the context of olive oil, 'virgin' as a premodifier has a strong emotional connotation that is unlikely to be neutral. This may make 'Virgin Threads' an appropriate trademark for a fashion brand; however, the meaning of 'untouched'/unexplored' is ambiguous in relation to clothing (environmentally-friendly, unworn and appropriate for the sexually-inexperienced are all possible interpretations). Tenor Plausibility is Medium.

**Mode**

**NOUN + VIRGIN**

The meaning of an NP + VIRGIN n-gram is heavily dependent on co-text. Most instances occur with a pre-modifying NP that is an anaphoric reference to the title or main topic of the article or blog post in a GIVEN-NEW organisation (e.g. “Pop Culture Trivia virgin”, “Zumba virgin”, “mood lighting virgin”). The NP is frequently a proper noun/name (e.g. "Tiger Tiger virgin"). Noun Phrases such as “buy-to-let virgin” or “FTSE-100-company-boss virgin” have no meaning outside of the texts in which they appear.

**VIRGIN + NP**

'Virgin' is an intertextual reference to external standards (legal definitions, product descriptions) when used to describe products such as olive oil, commercial hair and cocktails. Similarly, other uses of 'virgin', e.g. to describe an inexperienced person or environmental features, are commonly factual assertions.

**Plausibility Evaluation**

It is possible that the average consumer could be aware of a verifiable standard of 'virgin' in relation to clothing materials, although she would need to be interested in market innovations and trends. Consequently, Mode Plausibility for VIRGIN THREADS is Medium. There is little evidence for a construal of 'carbon' sufficiently defined to give meaning to the phrase 'carbon virgin' outside the context of a specific article, blog post or pun. Mode Plausibility for CARBON VIRGIN is Low.

**Discussion**

Judges' lexicographic analysis led to contradictory conclusions across three Carbon Virgin cases. Using dictionaries, the courts in the initial two cases agreed that 'virgin' could post-modify a noun phrase (when meaning “a person who is naïve or inexperienced on issues”) but corpus analysis suggests they failed to limit the range of
potential VIRGIN + NOUN/NP word combinations to proper or otherwise strongly referenced nouns. In the third case, although corpus analysis suggests his analysis of CARBON as an adjectival modifier comparable to the Oxford Dictionary example of 'political virgin' was overly restrictive, the judge was correct to conclude that CARBON VIRGIN has no clear meaning.

The issue of context was inconclusively addressed in two separate omnibus market research surveys commissioned by the legal parties in the Carbon Virgin dispute. Both offered evidence that the average consumer could associate CARBON VIRGIN with their business if given appropriate context; both were dismissed as evidence due to the use of leading questions and insufficiently representative samples. The Field and Mode parameters used in this analysis provide an alternative tool for evaluating the type and salience of context required to produce meaning.

Conclusions

Three findings emerge from this research:

1) The beneficial use of a multi-genre corpus confirms Butter's advice against using general corpora in trademark linguistics (given in Microsoft Corporation v. Apple Inc. 2010).
2) Register analysis can provide an empirical framework for inductive generation and evaluation of hypothetically plausible, 'suggestive' meanings and thus assist in determination of trademark protection.
3) Syntactic analysis demonstrates how the potential meaning of a trademark is influenced by word order - an under-researched area of trademark linguistics (cf. Heymann, 2010).

Two directions for future research are suggested: i) specification of Plausibility Ratings levels for computational linguistic analysis; ii) development of tools to assist legal professionals in the use of corpora in trademark linguistics.

References


**Legal Cases**

Abercrombie & Fitch Co. v. Hunting World, 537 F.2d 4, 2nd Cir. 1976
Microsoft Corporation. v. Apple Inc, Opposition No. 91195582, 2010
Michael Casey v. Virgin Enterprises Ltd, Opposition No. 98170, 2010
Virgin Enterprise Ltd v. Michael Casey, EWHC 1036, 2011
results/o37512.pdf
“It is important to reinforce the importance...”: ‘Hype’ in medical research articles

Neil Millar (University of Tsukuba, Japan), Francoise Salager Meyer (Universidad de los Andes, Merida, Venezuela) and Brian Budgell (Canadian Memorial Chiropractic College, Canada)

This paper assesses a phenomenon we call ‘hype’ – hyperbolic and/or subjective language to glamorise, promote and/or exaggerate in reporting of research. From a corpus of twenty-four randomized controlled trials (RCTs) in orthopaedic medicine we identified 161 hypes which we categorised for semantic target and linguistic realization. Hypes in RCTs are most prevalent in Discussion sections and most frequently serve to aggrandize the methodology. Findings are discussed in relation to factors including competition, pressure to publish in high-ranking journals and the influence of standardised guidelines.

1. Introduction

Research articles serve not just to inform but also to convince. Consequently, authors may be inclined to employ language which emphasizes, promotes or overstates what they see as positive aspects of their findings, their methodologies or even themselves. Vinkers et al. (2015) find that over the past 40 years there has been an increase in the use of overtly positive adjectives in scientific abstracts in PubMed (e.g. robust, novel and innovative). Similarly, Fraser and Martin (2009) show in medical articles an increase in the use of “value-laden words” to modify knowledge claims which might “bias” the reader’s interpretation. While such language may encourage the implementation of knowledge with beneficial clinical outcomes, it may also impose judgements on readers that undermine objective and disinterested evaluation of new knowledge, and, in the extreme, may represent a form of academic dishonesty.

Randomized Controlled Trials (RCTs) are type of study that aims to minimise bias when testing treatments – in medicine, they are generally regarded as the ‘gold standard’ of and can have a substantial influence on health care practice and policy (Cook et al. 1992). The purpose of this study is to provide quantitative and qualitative descriptions of how authors of RCTs use hyperbolic and/or subjective language to glamorise, promote and/or exaggerate aspects of their research – a phenomenon we refer to as ‘hype’

2. Corpus and Method

The following research questions were addressed: RQ1. To what extent do authors use hypes and how are hypes distributed across reports of RCTs? RQ2. What aspects of their research do authors choose to hype? RQ3. How are hypes linguistically realised?

Our corpus comprised 24 reports of RCTs with comparable subject matter from five leading journals in orthopaedic medicine (75, 927 tokens). Four articles were first independently manually annotated by each of the authors (two linguists,
one specialist in spinal research) for features that we considered as hype. These were then categorised according to their semantic target. Through multiple rounds of discussion, acceptable levels of agreement among analysts were reached. The framework for categorisation of targets is exemplified below (hype highlighted in bold).

Broad Research Area (BRA)

(1) **Accurate identification of the pain generators is critical**, particularly when ....

Specific Research Topic (SRT)

(2) **Compared with simple PKP, percutaneous internal fixation with PKP is a valuable surgical option** for ...

Authors' Prior Research (APR)

(3) Since [...] our original study in 2005 [...] and the subsequent implementation of our findings into practice ...

Research Methodology (RM)

(4) Patients eligible for inclusion were those undergoing primary total hip arthroplasty performed by the same trained surgeon (E.S.) using a direct anterior approach and

Research Outcome (RO)

(5) The latter benefits may be of **particular importance** for elderly patients, ...

Research Priorities (RP)

(6) The current study provides a notable contribution to the literature as being the first to demonstrate that ...

Research Conclusion (RC)

(7) The present study succeeded in demonstrating Level 1 evidence that locally administered depomedrol significantly decreases ...

Using this framework, each of the remaining twenty articles was then independently analysed by all three investigators. Respective analyses were reviewed and differences were resolved through discussion. Hypes, targets, information on linguistic realisations and article metadata were recorded in a spreadsheet.

3. Results and Discussion

We identified 161 hypes corresponding to 6.7 occurrences per-article, or 2.0 occurrences per 1000 words (henceforth ptw) (RQ1). The degree to which authors employ hypes varies considerably. Six articles contain just over 55% (89) of all hypes while two articles contain no instances. The distribution of hypes across sections and targets are shown in plots in Figure 1 (RQ1&2). Hypes are significantly more frequent in Discussions and Introductions (Fig 1A - overall frequency of 3.9 ptw, 2.8 ptw, respectively). Hypes targeting Research Methodology (RM) and Research Outcomes (RO) are significantly more frequent than other targets (Fig 1B). The variety of targets are greatest in the Introduction and Discussion sections (Fig 1C).
Corresponding to the typical IMRD ‘hourglass’, the distribution of hypes, both in terms of sections (1A) and targets (1C), appears to be consistent with the amount of rhetorical work we might expect authors to perform at different stages in a research article. That is, the Introduction serves to position the research, thus, requiring a wider range of rhetorical and thematic structures. Accordingly, authors in our sample employ a wide range of hypes. As the Methods and Results sections then narrow to focus on the description of the study, the use of hypes tapers and becomes fewer and more targeted. Finally, in the Discussion, as authors’ focus broadens to situate the research in the wider context, make claims and highlight strengths of the study, the use of hyperbolic language is at its highest level in terms of both the quantity and range of hypes – a bottom-heavy hourglass.

Our analyses indicate that hypes in RCTs are strongly associated with research methods. RM hypes with an animate target that serve to emphasise the expertise, qualifications and/or experience of individuals are limited to the Methods section (4, 8 & 9). Hypes in the Discussion tend to be more explicit, often stating a specific aspect of the study design as a ‘strength’ (10). In medical writing this strategy is not uncommon – indeed, BMJ guidelines recommend that authors devote two paragraphs to the ‘strengths and weaknesses’. We suggest both types of statement may be considered gratuitous. For example, it is normal that surgeons are trained, certified and expert (4, 8 &9), while the underlined elements in (10) highlight study design features which are requirements to qualify as an RCT, and, therefore standard.

(8) Assisted by 4 experienced nurse investigators (TMJ, LLD, AGM, and SMZ), a board-certified orthotist (CO) (RWT) measured the subjects to determine
the correct collar size and fitted the collars according to the manufacturers’ instructions.

(9) A primary diagnosis of degenerative or isthmic spondylolisthesis and/or LSS determined by expert spine surgeons.

(10) Strengths of this study are the introduction of a novel technique to the performance of ACDF with BMP-2 and its randomized, prospective, blinded, and controlled design ...

<table>
<thead>
<tr>
<th>Adjectives (n=76)</th>
<th>Adverbs (n=50)</th>
</tr>
</thead>
<tbody>
<tr>
<td>experienced</td>
<td>especially</td>
</tr>
<tr>
<td>comprehensive</td>
<td>greatly</td>
</tr>
<tr>
<td>fundamental</td>
<td>highly</td>
</tr>
<tr>
<td>specialized</td>
<td>most</td>
</tr>
<tr>
<td>relevant</td>
<td>importantly</td>
</tr>
<tr>
<td>senior</td>
<td>even</td>
</tr>
<tr>
<td>striking</td>
<td>particularly</td>
</tr>
<tr>
<td>tremendous</td>
<td>well</td>
</tr>
<tr>
<td>unwanted</td>
<td>obviously</td>
</tr>
<tr>
<td>appropriate</td>
<td>very</td>
</tr>
<tr>
<td>innovative</td>
<td>more</td>
</tr>
<tr>
<td>first</td>
<td>further</td>
</tr>
<tr>
<td>precise</td>
<td>markedly</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nouns (n=23)</th>
<th>Verbs (n=16)</th>
</tr>
</thead>
<tbody>
<tr>
<td>robustness</td>
<td>assure</td>
</tr>
<tr>
<td>experience</td>
<td>note</td>
</tr>
<tr>
<td>strength</td>
<td>succeed</td>
</tr>
<tr>
<td>importance</td>
<td>strengthen</td>
</tr>
<tr>
<td>maximize</td>
<td>recognize</td>
</tr>
</tbody>
</table>

Figure 2: Lexical realisation of hypes (size and shading proportional to frequency)

Linguistic categorisation of the hypes (RQ3) indicates that adjectives, the word class prototypically associated with evaluation (Hunston 2010), are the most frequent form by which hypes are realised (40%), followed by adverbs (23%). The word clouds in Figure 2 give an overview of how hypes are lexically realised. We discuss overlap with linguistic taxonomies of stance and engagement (Hyland 2005).

For example, attitude markers (e.g. important, significant, essential, critical), boosters (e.g. highly, greatly, entirely, obviously, it is not hard to see) and cognitive directives (e.g. A key issue to recognize ..., It is noteworthy that ...).

4. Conclusion

Although we find hype in RCTs, it is by no means uniformly prevalent. As a genre of study defined by a set methodology and viewed as the most rigorous study design
(Grossman & Mackenzie 2005), reporting of RCTs is highly constrained. All journals in our sample reference the ‘Uniform Requirements’ of the International Committee of Medical Journal Editors – these place constraints on format (e.g. IMRAD), content and ethical issues. In addition, all journals reference CONSORT, genre specific content guidelines. Such constraints are likely to restrict opportunity for author self-expression and lead to a genre that is quite formulaic.

Where hype does occur, it is most likely to be targeted at the methods. Given the prestige assigned to RCTs, we suggest that high frequency of hypes targeted at the methodology reflect authors’ perceived need to convince editors, reviewers and readers of the rigor of their research. The use of this strategy may, to some degree, be forced on the authors by constraints of the genre, standardised guidelines and journal policy (which at times are contradictory). That is, authors’ perceived need to highlight strengths may lead to the inclusion of unnecessary promotion. There is evidence that guidelines can influence authors’ linguistic choices, and do not necessarily lead to better reporting (Millar, Budgell & Fuller 2012). While it is beyond our scope to argue for or against the use of such promotional features, we suggest that these findings may inform journal guidelines for authors, editors and reviewers, as well as standardised guidelines.

References


Fraser, V.J. and J.G. Martin. (2009). Marketing data: has the rise of impact factor led to the fall of objective language in the scientific literature? Respiratory Research. 10: 35.


Assessing Text Difficulty Development in ELT Textbooks Series Using N-gram Language Models based on BNC
Alvin Cheng-Hsien Chen (National Taiwan Normal University, Taiwan)

This study aims to explore the possibilities of developing a corpus-based evaluation system of text difficulty development for English Language Teaching (ELT) materials by drawing insights from the N-gram-based language models used in natural language processing. ELT materials evaluation has paid little attention to the appropriateness of the text difficulty development in a textbook series (Ghorbani, 2011; Mukundan & Ahour, 2010; Tomlinson, 2012; Tsagari & Sifakis, 2014). The growing availability of machine-readable large corpora has facilitated the computation of frequency ratings for words, thus offering objective statistics for lexical sophistication in a given text (Crossley, Cobb, & McNamara, 2013). A heuristic in using the frequency ratings of words from large corpora is: words of lower frequency tend to be more “sophisticated”. Based on this principle, we aim to propose a new method for assessing the progression of lexical sophistication in ELT materials, or text difficulty in general, which will be argued to be a favorable measurement compared to the band-based frequency approach in the research paradigm of Lexical Frequency Profile (Laufer and Nation, 1995). Specifically, we addressed three research questions: (1) How can N-gram-based language models in NLP be applied to the field of computational text analysis and to the measurement of text difficulty? (2) Can an N-gram-based language model properly characterize the text difficulties development of an ELT textbook series and assess the developmental trends and transitional gaps of the text difficulty progression? (3) Does the N-gram-based method provide a better assessment of text difficulty development than the band-based frequency method in LFP?

For N-gram-based method, we built a trigram language model on British National Corpus (XML Edition), using self-developed scripts, by converting the raw frequencies of all the trigrams into probabilities via normalization. In case of unseen events, we adopted the state-of-art smoothing algorithm — Kneser-Ney Smoothing — to estimate the probabilities of the unseen (Jurafsky & Martin, 2008). Then the generated Trigram Language Model was used to compute the log probability of each sentence in a textbook volume, serving as a measure of degrees in lexical sophistication under the condition that the language model is representative enough. That is, each volume of the ELT textbook series was quantitatively measured by a series of log probabilities. As for the band-based LFP method, we used the BNC unlemmatized word lists provided by Adam Kilgariff as our base list and generated the top thirteen 1000-word lists according to the procedure described in Chen (2016). Following the LFP framework, the contents of each volume in the ELT textbook series was described by the coverage rates of each 1000-word list across thirteen frequency bands. Therefore, the text difficulty of a given text was represented by two different sets of metrics: Log Probabilities (N-gram method) and coverage rates (band-based method). Two experiments were conducted in this study.

Our first experiment adopted the bottom-up clustering-based method to identify the developmental trends and transitional gaps of text difficulty in a six-volume textbook series, using the algorithm of Variability Neighboring-joining Clustering, proposed by Gries and Hilpert (2008). We collected a corpus of senior high school textbooks (SHSTC), including three major officially-approved versions of textbook series used in the compulsory English curriculum in Taiwan. Each textbook series has six volumes targeted for use in six different semesters (in three years). The VNC proceeded as follows. The first step was to evaluate the distance between each pair of neighboring volumes (e.g. V1 vs. V2, V2 vs. V3, ..., V5...
vs. V6) and the second step was to construct a hierarchical tree by merging volumes of the minimal distance. In terms of the distance metric, we used the $p$-values from Kruskal Wallis Tests for the log probabilities from the N-gram method and Pearson's correlation coefficients for the coverage rates from the band-based method. In each step of merging, all the pairwise distance matrix was computed and the two neighboring volumes/super-clusters of the minimal distance were merged into one super-cluster. This merging proceeded iteratively until all the six volumes were merged into one node. After the dendrogram was created by the VNC, we created the plot of within-cluster variance (Kaufman & Rousseeuw, 2005) to determine the number of stages in the text difficulty development. The general principle is that the within-cluster variance decreases as the number of clusters increases. The local minimum (i.e. an elbow point) of this plot provides a possible optimal number of transitional gaps in our dataset.

In our second experiment, we evaluated the performance of the two metrics from N-gram and band-based methods, by applying them to analyze the text difficulty of four distinct types of ELT materials. In addition to the SHSTC, we further collected three types of ELT materials: (1) Junior High School Textbooks Corpus (JHSTC), (2) College Entrance Examination Corpus (CEEC), (3) Teacher Recruitment Examination Corpus (TREC). The JHSTC contained the official textbook materials used in the junior high school in Taiwan. The CEEC included all the entrance examination papers administered to the senior-high school graduates as an official qualifying exam for their college/university admission. The TREC included examination papers administered as qualifying exams for prospective English teacher recruitment in the secondary education in Taiwan. We assume that the difficulty levels should be: JHSTC < SHSTC < CEEC < TREC, where the junior-high school textbooks (i.e. JHSTC) are expected to be the least sophisticated texts and the teacher recruitment examinations are expected to be the most challenging texts from an L2 learner’s perspective. Under this assumption, we argue that a better metric of text difficulty should be the one that can maximize the difficulty levels among these four types of dataset. For each type of corpus, we computed their N-gram-based metric (i.e. log probabilities of sentences in each corpus) as well as band-based metric (i.e. LFP coverage rates in each corpus) as indices for text difficulty (or lexical sophistication). We then conducted a variance-based statistical test ($F$-test) to see which metrics would generate the maximal between-group variances. That is, the metric that can maximize the differences among these four corpora may better distinguish the variation of the lexical sophistication among these four text difficulty levels.

Our results show that the N-gram-based estimation of lexical sophistication yields comparable results to the prediction by the band-based method and the former generates a better metric that can maximize the variation of the text difficulties among the four types of ELT materials we collected. Furthermore, we argue that N-gram-based method is more favorable in several aspects. First, an N-gram language model does not have to make an arbitrary assumption with respect to the size of the frequency bands (e.g. 1000-word), thus running less risk of over-estimating the vocabulary level of the text. Second, the language model provides not only more fine-grained and precise estimates of probabilities for words but also probabilistic estimates for word combinations (e.g. multiword expressions), thus taking into account considerable degrees of contextual sophistication of collocation/colligation patterns. Third, the N-gram language model provides a more holistic measure for lexical sophistication by considering the low-frequency patterns in different frequency scales. Most importantly, these probabilities were based on a large-scale representative corpus, which provides a common ground for cross-cultural comparative evaluations of ELT materials, thus shedding light on several pedagogical implications for EFL/ESL learners and teachers as well as ELT textbook developers.
References


Fang Wang (University of Essex, UK)

1. Introduction: research background and research questions

This research falls within the broad scope of corpus-based discourse analysis. It looks at a specific social phenomenon that is increasingly relevant to human life: drug advertising. Advertising language has been extensively investigated in the field of discourse analysis. Fairclough (1993) examines the advertisements of academic posts and identifies the promotional element in the newer university, explaining that the marketization of universities is the wider social practice within which these discourse practices underwent significant change. Koteyko & Nerlich (2007) investigate the discourse of probiotic web advertising, showing how appeals to shared values together with associations with high-tech science function in generating trust and reinforcing the discourse of healthism. Baker (2006) presents a critical discourse analysis of holiday advertising leaflets: by using corpus linguistic methods, he depicts the most frequently employed language patterns and identifies the social matrix and over-arching ideology related to tourism constructed in the advertising leaflets. This research aims to, by using corpus methods, identify the linguistic patterns of antidepressant advertisements in China, and explore whether such patterns have changed/developed for the last 20 years. Furthermore, it attempts to explain how these patterns and changes are related to their wider patterns of sociocultural contexts.

2. Data collection and management

The special corpus built for this research includes all the antidepressant advertisements that appear in the Chinese Journal of Psychiatry from 1996 to 2015. Corpus research methods such as frequency, collocation, keyword and concordance analyses are conducted by using the WordSmith Tool (5.0). In analysing Chinese corpora, the ICTCLAS (the Institute of Computing Technology, Chinese Lexical Analysis System) is be used to segment Chinese texts into analysable lexical units. The reference corpus used for keyword analysis is the Lancaster Corpus of Mandarin Chinese (LCMC). To highlight the distinctive features of the linguistic patterns of Chinese adverts, 10 adverts randomly chosen from the British Journal of Psychiatry are included for a comparative analysis.

3. Corpus analyses and main findings

The findings of frequency analysis show that the word 焦虑 (anxiety) is the most frequent word in the Chinese corpus, with 治疗 (treat) as the second. Concordance lines of 焦虑 (anxiety) show clearly that this word is predominantly used to describe the symptom that antidepressants can treat. Other symptoms that antidepressants treat are heavily simplified, such as by using the phrases of depressive feeling and all types of depression. In British adverts, the use of a particular brand of
antidepressant is described in relation to specific types of depression with distinctive symptomatic features, providing more detailed information for prescribing doctors.

Collocate analysis enables to see all the words that co-occur with (adverse) effects, and the top five collocates are comparatively, and, common, with and include. Concordance lines in which both adverse effects and these collocates occur show that the side-effects of antidepressants are represented as comparatively rare, and the common side-effects of antidepressants include very light stomach discomfort, etc. By contract, the category of adverse effects in British adverts offers very clear and straightforward contents which are identical with the information included in medical instructions.

A diachronic keyword analysis indicates that from 1996 to 2002, two important key words appeared as effectiveness and effective, implying that antidepressants are constructive as effective in treating depressive symptoms, such as anxiety and depressive feeling, and such effectiveness is repeatedly emphasized. From 2003 to 2008, more formal and scientific names of the categories of information start to appear, such as indication, and adverse effects, implying that Chinese adverts began to develop in a more scientific and objective way. From 2009 to 2015, contraindication and hypersensitivity entered as new key words, indicating that Chinese adverts in the last ten years have begun to include information related to the contraindications of antidepressants.

4. Discussion and implications

Antidepressants are represented as very effective, with rare side-effects, and convenient-to-use medication in Chinese adverts, with rather general and simple languages. Despite an increase of the frequency of relevant adverts in medical journals and an emphasize in adding more balanced information in adverts, antidepressants remain a less discussed medication in Chinese society. This is closely related to the ways how depression and antidepressants have been conceptualized in Chinese culture. Wang (2013) finds that China constructs ‘抑郁症’ (yiyuzheng, ‘depression’) as a problem that is normally caused by external social factors, and therefore psychological support and improvement of the social environment have been represented as more helpful. Medical treatment, on the other hand have been marginalised and negatively represented. By contrast, Britain has constructed depression as both a psychological problem that needs psychotherapy and a biochemical condition that needs pharmaceutical intervention. Especially in the last two decades, with the rapid development of pharmaceutical industry in the UK, depression has been predominantly constructed as a chemical condition and over 50 million prescriptions of antidepressants have been written for the British. The antidepressant advertising regulations have been established and the existing policy structure has ensured that more balanced and truthful information of antidepressants can be represented in the UK. This research, therefore, demonstrated the view that all the factors beyond linguistic structures, such as those in the context of situation and context of culture, are intricately woven into the actual use of language. Even drug advertising, which is perceived as an objective form of language, can be devised in dramatically different ways in Chinese and British societies. Discourse, therefore, can be understood as the totality of all the
texts that have been produced within a particular discourse community (Teubert, 2010). By discovering the features and patterns in different discourses, one can always find how a specific discourse object has been differently constructed and understood. This research is an integrated part of a more large-scale project, comparing the constructions of antidepressants in British and Chinese medical advertisements, in national newspapers and in medical journals. It will increase the reflective awareness of both laypeople and health professionals as to how news and scientific discourses echo, filter or reject the assumptions made in consumer discourse, showing that medicine is not an objective and self-contained entity, but that medical conditions and their implications are discursively constructed.

References

Discourse practice and news online: comparing discursive constructions of Romanian immigrants in articles published by the Daily Express with readers’ comments
Mark McGlashan (Birmingham City University, UK) and Paul Baker (Lancaster University, UK)

Introduction

This paper presents results of research into the representation of Romanian immigrants in online news articles, and readers’ comments on those articles from a corpus-assisted critical discourse studies perspective. Focus on European immigration into the UK (especially Eastern European countries such as Romania) is a timely political issue in the UK where nationalistic, anti-immigration political parties such as the UK Independence Party (UKIP) have grown rapidly in size, votership, and influence – the 2015 UK general election saw UKIP secure 12.6% of votes (a rise of 9.5% compared to the previous election, and the highest rise of any party). More recently, on 23 June 2016, a national referendum was held across the UK to decide whether the country should ‘Remain’ in or ‘Leave’ the European Union, popularly referred to as Brexit (‘British Exit’). 51.89% of the votership voted ‘Leave’ as opposed to 48.11% voting ‘Remain’. Following the vote, a poll of over 12,000 people carried out by Lord Ashcroft gave the top three reasons for voting Leave as ‘decisions about the UK should be taken in the UK’, ‘the best chance for the UK to gain control over immigration and its borders’ and ‘remaining meant little or no choice about how the EU expanded its membership or powers’. Regarding Romania specifically, Romania joined the European Union in 2007 but a transitional cap on migration meant that Romanians were not able to become resident in the UK until the beginning of 2014 (unless they were self-employed or worked in seasonal jobs).

Critical Discourse Analysis (CDA) (which is being increasingly combined with Corpus Linguistic (CL) approaches (Baker and McEnery, 2015)) understands discourse as being socially constructive and constitutive; discourse “constitutes situations, objects of knowledge, and the social identities of and relationships between people and groups of people” (Fairclough and Wodak, 1997: 258). As such, much research in the field has focussed on the representations and discursive constructions of identities including, for example, gender and sexuality (Potts, 2015, Turner, 2015). And given that a central interest in CDA concerns the critique of power relations and ideology (especially those that are inequitable; Wodak and Meyer, 2009), much of this work has sought to interrogate those constructions that create or reinforce states of social inequity with regards to, inter alia, sexism (Lazar, 2005), xenophobia (Baker et al., 2013), and racism (Reisigl and Wodak, 2001). Our focus here on Romanian immigrants contributes to this body of literature and extends current understanding about discourses surrounding RASIM (Refugees, Asylum Seekers, and Immigrants) in the UK press (Gabrielatos and Baker, 2008, KhosraviNik, 2010, KhosraviNik, 2009).

Concerning the press, news media is a site of particular interest for CDA researchers due to the powerful social position it holds (Fairclough, 1995). Scholars recognise that journalistic discourse can be socially constitutive and transformative “through shaping understandings, influencing audience attitudes and beliefs (particularly through their reinforcement), and transforming the consciousness of those who read and consume it” (Richardson, 2007: 29). However, the idea that readers uncritically consume news media has been roundly problematized by CDA researchers. Fowler, for example,
argues that readers are not simply “passive vessels or sponges, absorbing an ideology which the source of the text imposed on them” but that “[t]here is every reason to propose that being a reader is an active, creative practice” (1991: 43). Indeed, following a shift to the distribution of news online, the role of the reader in the construction of news events is being given greater attention (McEnery et al., 2015) and O’Keeffe notes that “the reader is no longer reading an article in protracted isolation; s/he can comment on it via a website, email it to a friend, post it on a social network for others to discuss it” (O’Keeffe, 2011: 450).

Recognising this, our paper focuses on this changing relationship between text and ‘discourse practice’ – i.e. the “various aspects of the processes of text production and consumption” (Fairclough, 1995: 58; figure 1) – and attempts to address Richardson’s claim that “this aspect of CDA remains the most under-developed” (Richardson, 2007: 39). We suggest that CDA has been slow to address this underdevelopment largely due to a lack of available methods for data collection rather than deficit of theoretical or analytical approaches, and so present a method for constructing corpora of online news texts and their readers’ comments which can then be analysed to examine and compare the relationships between the content of the news articles and readers’ comments on (i.e. consumer responses to) those articles. We suggest that comments on online news articles act as type of proxy for reader reception/consumption and thus fit well with CDA’s understanding of the contextual ‘situatedness’ of a text.

Research questions

As this study is interested in discourses around Romanians in articles printed by the Daily Express, our first research question is:

1. How were Romanians typically represented by the Daily Express in the time leading up to Brexit?

Moreover, because of our further interest in comparing discourses found in both articles and readers’ comments on the online versions of those articles and what this tells us about discourse practice, our second research question is:
2. What were the differences/similarities in representation between articles and reader comments?

**Sampling & data**

In order to focus on both news articles and readers’ comments on those articles, we collected both forms of data by building a ‘web scraping’ tool using Python and the Python libraries BeautifulSoup and Selenium to extract all text in the articles and comments of pages containing the term ‘romanian’ published by the Daily Express online between 24th July 2006 and 23rd June 2016 (the date of the Brexit vote).

These sampling criteria returned 2 discreet corpora: a 771,878-word articles corpus containing 1,945 texts, and a 2,166,148-word comments corpus. A timeseries analysis (Figure 2) showed an increase in the number of texts in the articles corpus during – and especially towards the latter part of – 2013 when the EU lifted restrictions on Romanian and Bulgarian workers’ ability to take up residency in the UK.

![Articles published per day (pre-Brexit) containing 'romanian'](image)

*Figure 2: timeseries analysis*

**Methods, initial findings & further work**

Keyword and word-cluster analyses were employed to identify similarities and differences between the articles and comments corpora. During keyword analysis the comments and articles corpora were compared to the BE06 – a 1 million word reference corpus of written British English (Baker, 2009) – and, concentrating on the 50 most statistically significant keywords, found that several keywords were shared between the corpora. These shared keywords related to:

- politics (*UKIP, Cameron, Labour, Farage*)
- nationality (*EU, Romania, Romanian, Romanians, Britain, British, UK, Europe, countries*)
movement of people (migrants, immigration, immigrants)
- other unclassified terms (will, benefits, racist)

One dominant theme in the corpora found through keyword analysis was a focus on the quantification of Romanians and focus on their movement (e.g. ‘It estimates that about 50,000 Romanians and Bulgarians will arrive on our shores’). This finding is consistent with previous research on the representations of immigrants (Gabrielatos and Baker, 2008).

Initial word cluster analysis was performed by producing a list of 3-word clusters for the articles and comments corpora both, which were then intersected to produce a list of shared clusters. Numerous clusters including the keyword UKIP were identified as being shared between the articles and comments corpora and an initial investigation focussed on frequently occurring clusters which contained grammatically open-class words (e.g. 'ukip leader nigel', 'farage and ukip', 'support for ukip') but not those containing closed-class words (e.g. 'ukip in the', 'to the ukip'). One cluster – 'attack on ukip' – was found to occur in both the comments and the article relating to a page entitled, "It's the politics of anger' Cameron launches attack on Ukip'. The article focuses on responses to a statement made by Nigel Farage that he would be uncomfortable if Romanians became his neighbours. The cluster 'attack on ukip' is found in the headline for this article but also in a comment that appears to respond directly – and critically – to this headline. The comment begins:

1) Well apparently according to the Daily Fail Cameron has launched yet another attack on UKIP and the people choosing to vote for it. Calling us party of angry politics

[...]

The comment goes on to distinguish UKIP from mainstream political parties, creating in/out groups:

2) Cameron, Clegg etc have all taken turns to accuse British patriots and ordinary people of being racists, bigots, loons, fruit cakes etc and I am sure none of those big three parties would want bigots, racists and fruitcakes voting for them would they?

The commenter continues by imploring that readers vote for UKIP – something not found in the article. Results suggest that by combining keyword and word cluster analysis this method enables a corpus-driven ‘way in’ to quantitative and qualitative analyses at the level of the text as well as discourse practice. The work also demonstrates how the building ad hoc web scraping tools for specialised corpus creation can benefit CL research/researchers.

References


Guided reading: Using corpus methods to investigate how teacher strategies differ across children’s reading ability, SES, and teacher experience
Liam Blything, Kate Cain and Andrew Hardie (Lancaster University, UK)

1. Introduction

The quality of child-directed input has a strong impact on children’s learning and language development (Hoff, 2003; Vygotsky, 1978). Crucially, this is most influential when combined with methods that facilitate interaction (e.g., Pianta, Hamre, & Allen, 2007). Such an approach is widely implemented within the literacy classroom through guided reading (Oxford Primary, 2014). Although previous research has demonstrated the effectiveness of many of the techniques used in guided reading (e.g., Taylor, Pearson, Clark, & Walpole, 2000), there is substantial variability in what teachers do and, therefore, in our understanding of the specific strategies that should be used to apply those techniques (Ford, 2015). This paper aims to use large-scale corpus methods in order to identify the fine-grained language features of teacher input, and to examine how these features vary according to children’s age, reading ability, SES, and teacher experience.

2. Background

In guided reading sessions, a teacher works together with a small group of between 4 to 7 children on a reading activity and adapts their input according to learner responsiveness (Fountas & Pinnell, 1996; Oxford Primary, 2014). A teacher is expected to progressively increase the challenge of the session for children who display higher levels of competence in order to encourage a more independent and adult like contribution. This technique is used to scaffold the discussion and provides opportunities for children to engage in higher-level thinking about the text. For example, the teacher might ask more demanding questions or comments, or remodel a child’s utterance into a more meaningful and grammatically advanced utterance.

Previous research has demonstrated the effectiveness of many of the techniques used in guided reading. For example, Taylor et al. (2000) reported that children who displayed higher growth in reading assessments over a 5 month period, were more likely to have teachers who provided small group instruction (vs. whole group), practiced more scaffolding of the discussion (vs. telling), and asked more higher-level questions (vs. low level). However, surveys indicate that teachers lack awareness for the specific strategies that should be used in guided reading (Ford, 2015; Ford & Opitz, 2008). Moreover, observations indicate that sessions are still dominated by teacher talk (i.e., telling), despite the recommendation that teachers should place emphasis on scaffolding the discussion (e.g., Frey & Fletcher, 2010).

The findings outlined above motivate a more detailed understanding for how guided reading is implemented. Previous research on this topic has typically taken the form of observational studies, in which researchers have had to laboriously parse and hand-code transcriptions of the teacher–children interactions (a corpus) to
identify teacher strategies of interest. Because this is a long and painstaking process, it limits the size of the corpus to one that can be coded within a realistic time window. In this project, we maximise interpretation of naturalistic classroom interactions by using powerful corpus search tools. These enable precise computer searches for a wide range of language features, and are much faster and more reliable compared to hand-coding. This affords us to create and explore a much larger corpus of guided reading sessions than in previous studies, making a fine-grained analysis possible.

Our first aim is to identify the fine-grained language features of teacher input. Second, we examine how these features vary according to children’s age, reading ability, SES, and teacher experience.

3. Method

We collected a corpus of teacher-child interactions during guided reading sessions, approaching 500,000 words. The corpus represents recordings of 103 sessions that each typically lasted between 15 and 35 minutes. Recordings aimed to be non-invasive to the normal proceedings of a session. Teachers (N = 22) and children (N = 119; aged 5 to 11 years) were from school districts serving different socio-economic status. The language features of teacher input are analysed using the corpus tool CQPweb (Hardie, 2012). This process draws concurrent work that successfully created a set of corpus queries which can identify a range of question types that may be involved in scaffolding language (Smith, 2016).

Background measures were taken of children’s reading ability, and teacher experience.

4. Results and discussion

Early exploration of the corpus indicates that we will uncover a wide range of language features that can be organised onto a continuum from low challenge strategies (e.g., asking wh-questions that constrain responses, such as what, who and where) to high challenge strategies (e.g., asking wh-word questions that elicit evaluative responses, such as how and why).

We anticipate that low challenge questions and strategies will be more frequent for younger age groups and for groups with lower reading ability and that high challenge questions will be more common for older age groups and groups with better reading skills. This can be attributed to the notion that 4-year-olds are only just beginning to develop their ability to read words on a page, so teachers are more likely to focus on ways of improving their accurate translation of print into word meanings (i.e., decoding skills and vocabulary). Conversely, older children are able to read words relatively well, so teachers are more likely to target improving an understanding of the language that has been accessed from the printed word (i.e., reading comprehension skills).

In addition, the data is explored in relation to SES and teacher experience, particularly with regard to how these factors influence the trade-off between teacher-dominated ‘telling’ talk versus language scaffolding. The findings will also provide the groundwork for a future investigation into how language features might influence the quality of children’s responses.
References


This paper introduces a new measure of corpus similarity called SMF (Sum of Minimum Frequencies). This measure is conceived as an expected proportion of shared words in two large samples with replacement drawn from the corpora under comparison. SMF is shown to achieve comparable results to other word frequency measures of corpus similarity, namely, Spearman and $\chi^2$. The paper also discusses some possible modifications of the SMF metric involving trimming vocabulary lists and applying a power function to corpus frequencies.

1. Introduction

The problem of what is corpus similarity and how to measure it has been widely discussed within corpus linguistics, starting with Kilgarriff (1997) and Kilgarriff & Rose (1998). Since then, many measures of corpus similarity have been proposed (see Kilgarriff 2001, Kilgarriff 2009, Fothergill, Cook, & Baldwin 2016 among others), based on word frequencies, perplexity, and topic modelling. However, no definitive measure for corpus similarity has yet been found. In this paper, I propose and evaluate a new corpus similarity measure based on word frequencies. This measure is easy both to compute and to interpret.

2. Sum of Minimum Frequencies (SMF)

Let us assume that we have two corpora $C_1$ and $C_2$ we want to compare to each other. These corpora use a vocabulary $W = \{w_1, w_2, ..., w_n\}$. The relative frequencies of a word $w$ in these two corpora are $f_1(w)$ and $f_2(w)$. Now let us take two samples with replacement $S_1$ and $S_2$ from $C_1$ and $C_2$, respectively. If $S_1$ and $S_2$ contain a large proportion of shared words, $C_1$ and $C_2$ can be considered similar, and if this proportion is small, the corpora are dissimilar.

For instance, let us take three corpora $C_1$, $C_2$, and $C_3$ with a union of their vocabularies comprising five words $\{a, b, c, d, e\}$. The frequencies of these words in the three corpora are given in Table 1:

<table>
<thead>
<tr>
<th></th>
<th>$C_1$</th>
<th>$C_2$</th>
<th>$C_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$w_1 = a$</td>
<td>0.4</td>
<td>0.5</td>
<td>0.2</td>
</tr>
<tr>
<td>$w_2 = b$</td>
<td>0.3</td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td>$w_3 = c$</td>
<td>0.2</td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td>$w_4 = d$</td>
<td>0.1</td>
<td>0.0</td>
<td>0.1</td>
</tr>
<tr>
<td>$w_5 = e$</td>
<td>0.0</td>
<td>0.1</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Table 1. Three corpora for comparison: a toy example

Samples of 100 words drawn with replacement from each corpus would look like the following:
Samples $S_1$ and $S_2$ have 70% of words in common (37 $a$'s, 16 $b$'s, and 17 $c$'s out of 100 words), $S_1$ and $S_3$ have 44% of words in common (20 $a$'s, 6 $b$'s, 7 $c$'s, and 11 $d$'s), as well as $S_2$ and $S_3$ (20 $a$'s, 6 $b$'s, 7 $c$'s, and 11 $e$'s). This shows that $C_1$ and $C_2$ are more similar to each other than any of them is to $C_3$, which is the conclusion we would expect looking at the frequency lists.

In order to calculate the similarity score, we do not actually need to draw random samples, since frequency lists provide sufficient information to estimate the expected value of the sample intersection size. It is simply the **Sum of Minimum Frequencies (SMF)** for all the words in the vocabulary:

$$SMF = \sum_{i=1}^{n} \min(f_1(w_i), f_2(w_i))$$

Since $f_i(w)$ are relative frequencies by definition, SMF can range between 0 and 1, where 0 stands for absolute dissimilarity (it means that the two corpora do not share a single vocabulary item), and 1 stands for complete identity of the frequency lists for the two corpora.

To my knowledge, such a method was first used in biodiversity studies (Renkonen 1938). Drawing random samples and comparing them was proposed by Labbé & Labbé (2012). Shaikevich (2015) also compares corpora along similar lines, but he takes only $k$ most frequent words from the corpora for comparison.

### 3. Evaluation

The standard way to evaluate a similarity metric is to use Kilgarriff’s (2001) Known-Similarity Corpora (KSC) approach. In order to evaluate the SFM metric, I ran a series of experiments on the same set of BNC subcorpora as Kilgarriff had done:

Accountancy (acc); The Art Newspaper (art); British Medical Journal (bmj); Environment Digest (env); The Guardian (gua); The Scotsman (sco); Today (tod).

For each pair of these seven corpora, a KSC-set comprising 11 corpora was generated (e.g., 0% acc and 100% gua, 10% acc and 90% gua, ..., and 100% acc and 0% gua, no text fragment appearing in more than one corpus), and for these 21 KSC-sets all 660 KSC similarity judgments were checked against the a priori gold standard judgments. The percentage of times the SMF measure agreed with the gold standard is shown in Table 2, the mean percentage of agreement being 93.57 and the median being 96.21:

<table>
<thead>
<tr>
<th></th>
<th>acc</th>
<th>art</th>
<th>bmj</th>
<th>env</th>
<th>gua</th>
<th>sco</th>
<th>tod</th>
</tr>
</thead>
<tbody>
<tr>
<td>acc</td>
<td>94.39</td>
<td>90.76</td>
<td>98.33</td>
<td><strong>85.00</strong></td>
<td>75.76</td>
<td>92.27</td>
<td></td>
</tr>
<tr>
<td>art</td>
<td>96.52</td>
<td>99.70</td>
<td><strong>95.45</strong></td>
<td>89.85</td>
<td>98.03</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

Page 310
Table 2. The percentages of KSC gold standard judgments correctly captured by SMF

<table>
<thead>
<tr>
<th></th>
<th>bmj</th>
<th>env</th>
<th>gua</th>
<th>sco</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>99.24</td>
<td>96.06</td>
<td>97.88</td>
<td>98.03</td>
</tr>
<tr>
<td></td>
<td>99.24</td>
<td>98.48</td>
<td>99.39</td>
<td></td>
</tr>
<tr>
<td></td>
<td>80.30</td>
<td>95.30</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>84.85</td>
</tr>
</tbody>
</table>

Table 3. The percentages of KSC gold standard judgments correctly captured by five corpus similarity measures, reproduced from Kilgarriff (2001: 127)

<table>
<thead>
<tr>
<th>KSC-set</th>
<th>spear</th>
<th>$\chi^2$</th>
<th>closed</th>
<th>type 1</th>
<th>type 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>acc_gua</td>
<td>93.33</td>
<td>91.33</td>
<td>82.22</td>
<td>81.11</td>
<td>80.44</td>
</tr>
<tr>
<td>art_gua</td>
<td>95.60</td>
<td>93.03</td>
<td>84.00</td>
<td>83.77</td>
<td>84.00</td>
</tr>
<tr>
<td>bmj_gua</td>
<td>95.57</td>
<td>97.27</td>
<td>88.77</td>
<td>89.11</td>
<td>88.77</td>
</tr>
<tr>
<td>env_gua</td>
<td>99.65</td>
<td>99.31</td>
<td>87.07</td>
<td>84.35</td>
<td>86.73</td>
</tr>
</tbody>
</table>

SMF outperforms the three perplexity measures ("closed", "type 1", and "type 2") in all four test cases. It comes extremely close to Spearman ("spear") and $\chi^2$ in three out of four test cases, outperforming Spearman on the bmj_gua set and outperforming $\chi^2$ on the art_gua set. This shows that the SMF metric achieves results comparable to the corpus similarity measures that were deemed best by Kilgarriff (2001).

4. Modifications of the SMF metric

A possible drawback of SMF is that it gives either too much or too little weight to high-frequency items, such as articles, pronouns, modal verbs, etc. In order to account for this, I apply different types of transformations to test whether increasing the weight of high- or low-frequency items would provide a performance gain.

One option to think of is to trim the frequency distributions from above, i.e. to exclude the union of the words that occupy top $k$ positions in each corpus compared, thus removing some number of words between $k$ to $2k$ from the vocabulary. However, it turns out that trimming from above does not bring any systematic improvement. The median percentage of agreement with the gold standard for the same 21 KSC-sets ranges between 95.30 and 96.36 (see Figure 1 below), but this is definitely not enough to say that trimming some positive number $k$ of words from above is preferable to non-trimming, which yields a median of 96.21%:
Another possible option is to apply a power function to corpus frequencies. From the original frequency distribution \( \{ f(w_1), f(w_2), \ldots, f(w_n) \} \) we get the transformed distribution \( \{ f'(w_1), f'(w_2), \ldots, f'(w_n) \} \) using the following formula:

\[
f'(w_i) = \frac{f^p(w_i)}{\sum_{j=1}^{n} f^p(w_j)}
\]

Setting \( p \) below 1 increases the weight of low-frequency items (in the extreme case where \( p = 0 \), all words in a corpus have equal impact on the outcome regardless of their frequency), whereas setting \( p \) above 1 increases the weight of high-frequency items. Testing different values of \( p \) between 0 and 2 with a step of 0.1 shows that the median for our 21 KSC-sets is the same at \( p = 1.0, 1.1, 1.2, \) and 1.3 and equals 96.21% (see Figure 2).

Interestingly, setting \( p \) to 1.1 or to 1.2 improves the performance on the four KSC-sets that are directly comparable to Kilgarriff’s (2001) results, so that SMF even outperforms both Spearman and \( \chi^2 \) two times out of four.

Another possible way of improving the performance of the SMF metric might be trimming the data from below, i.e. leaving only the union of \( k \) most frequent words from both corpora for comparison, thus neglecting the infrequent items. This method was first proposed by Shaikevich (2015). Let \( V \) be the union of \( k \) most frequent words from both corpora. The set \( V = \{ v_1, v_2, \ldots, v_m \} \) contains some number of words \( m \) between \( k \) and \( 2k \), and the formula for this version of SMF is as follows:
\[ SMF(k \text{ most frequent}) = \frac{\sum \min(f_1(v_i), f_2(v_i))}{\sum \max(f_1(v_i), f_2(v_i))} \]

This tweak makes the SMF even more accurate. Figure 3 shows that the best results are achieved with \( k \) around 1000:

![Figure 3. SMF with trimming from below (i.e. leaving only the union of \( k \) most frequent words from both corpora)](image)

However, the performance gain achieved by the modified versions of the SMF metric might be due to the nature of the test data. The effect of various tweaks on the accuracy of the SMF has to be tested more thoroughly. For this reason, it is sensible to stick to the simplest version of SMF for the time being, since it also has the advantage of being easily interpretable in a way described in Section 2.

5. Conclusion

The Sum of Minimum Frequencies (SMF) metric presented in this paper is yet another corpus similarity measure based on word frequencies. It outperforms perplexity measures and achieves results comparable to those of Spearman’s rank correlation coefficient and \( \chi^2 \). Some modifications applied to this metric may eventually lead to SMF outperforming the other two measures. However, even the simplest version of the SMF, outlined in Section 2, performs quite well and has an advantage of being easily interpretable as a proportion of shared words in two large samples with replacement drawn from the corpora under comparison.

Further directions of research include a more in-depth study of the modified versions of SMF, confronting SMF with the keyword-based corpus similarity measure outlined in Kilgarriff (2009) and implemented in SketchEngine, and comparing SMF to other corpus similarity measures using languages other than English.

References


Most research on lexical change has analysed variation in the use of relatively few words over relatively long periods of time (Traugott and Dasher, 2001). In part this is because of limitations in the quantity of language data that has been available for analysis. As most words are very rare (Zipf, 1949), it is difficult to study change in word usage on a large scale based on small corpora. The situation, however, has recently changed with the rise of online communication, especially social media, which has allowed for extremely large and densely sampled corpora to be compiled. These new multi-billion word corpora have opened up several areas of research on lexical variation and change (e.g. Bamman et al., 2014; Eisenstein et al., 2014), including making it possible to analyse change in the relative frequencies of large numbers of words simultaneously over very short periods of time. To better understand the mechanisms of short-term lexical change in modern English, this paper presents a multivariate analysis of change over time in the relative frequencies of tens of thousands of word forms based on a multi-billion word time-stamped corpus of American social media data from 2013-2014. Rather than analyse change in the frequency of each word individually, this set of words was analysed together using a multivariate statistical analysis to identify the most common time series patterns in this variety, allowing for the dynamics of lexical change to be analysed from a new perspective and in unprecedented detail.

This analysis is based on an 8.9 billion word corpus of American Tweets collected between October 2013 and November 2014 using the Twitter API (see Huang et al., 2016; Grieve et al., 2017; Nini et al., 2017). In order to analyse general patterns of lexical frequency change in this corpus, the relative frequency of the 67,022 most common words in the corpus (all words forms that occur at least 1,000 times in the complete corpus) were measured across the 397 days in the corpus yielding 67,022 relative frequency time series. Each lexical time series can be visualised as a line graph showing how the relative frequency of that word form has changed over the days in the corpus. This complete dataset was then subjected to multivariate statistical analysis to identify common patterns of diachronic variation.

The time series for each word was first scaled to range between 0 and 1 to preserve the shape of the time series while allowing for the set complete set of words to be compared to each other despite occurring across vastly different relative frequency ranges. The 67,022 scaled time series were then subjected to a non-linear iterative partial least squares (NIPALS) principal component analysis (PCA). A PCA is a common multivariate statistical method (Everitt & Hothorn, 2011) that takes a set of variables (in this case, word relative frequencies) measured over a set of observations (in this case, days of the corpus) and extracts a series of principal components that represent the most important patterns of variation in that dataset (in this case, time series). In other words, PCA is a dimension reduction technique: starting with a large set of variables that generally exhibit substantial inter-correlations, PCA identifies a smaller set of uncorrelated dimensions that maximally explain the variability exhibited in the original set of variables. A NIPALS PCA (Wold, 1975) was used in this study because it is a version of PCA that allows for very large
multivariate datasets to be processed efficiently, including datasets with far more variables than observations, as is the case here, which is not possible using a standard PCA. Based on the NIPALS PCA of the scaled time series, the most important patterns of lexical frequency change in the dataset were identified, allowing both for these patterns to be visualised (by plotting the component scores) and for the degree to which the time series for each of the individual words is characterized by these aggregate patterns to be assessed (by inspecting the component loadings).

The first 10 dimensions identified by the NIPALS PCA are plotted in Figure 1, ranked in order of importance (i.e. by amount of variance explained). Each of these graphs plots the component scores for one dimension, in essence allowing for the 10 most common patterns of relative frequency change over time in the 67,022 word forms in the dataset to be visualised. In each case, two time series are plotted on the graph, which are mirror images of each other. This is because each dimension extracted by the PCA identifies words that not only show highly positively correlated patterns but also words that show highly negatively correlated (i.e. complementary) patterns. For example, the first dimension identifies words whose relative frequencies show a relatively consistent rise or fall over time. A secondary weekly trend is also visible in Dimension 1, with rising words tending to be used especially frequently on Fridays and Saturdays. In addition to plotting the general shape of these trends, inspecting the component loadings allows for word forms that most clearly exhibit that pattern to be identified. In the case of Dimension 1, rising words include various proper nouns (e.g. *timehop*, *soliant*) and recent word formations (e.g. *fleek*, *fav*), whereas falling words include numerous interjections (e.g. *haha*, *uh*). Subsequent dimensions can be interpreted in similar ways, including identifying both smaller and larger cyclical trends. For example, Dimension 2 primarily identifies a weekly pattern, with certain words being more common on weekends (e.g. *hangout*, *fun*) or weekdays (e.g. *desk*, *lecture*), although a secondary seasonal pattern is also apparent, with weekend words being more common in the summer and weekday words in the winter. Similarly, Dimension 3 identifies a more focused distinction between summer (e.g. *vacation*, *popsicles*) and winter (e.g. *butternut*, *jacket*) words.

Identifying common temporal trends in the relative frequencies of a large numbers of words can inform our understanding of lexical change in numerous reasons. For example, this analysis finds that rising words generally follow a roughly s-shape curve of change on Twitter, which agrees with standard theories of linguistic innovation (e.g. Blythe & Croft, 2012). This analysis also finds that rising words tend to be used especially frequently on weekends, suggesting that certain topics and people may be more or less likely to drive lexical change, at least in this variety of language. More generally, most dimensions show both short-term (e.g. weekly) and long-term (e.g. monthly) cycles, demonstrating that word frequencies pattern at multiple temporal scales concurrently. The long-term cycles that characterise all but the first dimensions appear to primarily reflect cyclical topical trends, raising questions about how such trends might affect the rate of lexical change in different semantic domains.
Figure 1  First 10 PCA Lexical Time Series Dimensions
References


Corpora in teacher training through Pedagogical Grammar –
A case study of two workshops
Balázs Vida (Eötvös Loránd University, Hungary)

Background

As several surveys conducted with practising teachers of English revealed (see Mukherjee, 2004; Römer, 2009; Tribble, 2015), even if corpora are present in language education, direct uses of corpora in secondary, let alone primary, school settings are still uncommon (McEnery & Xiao, 2010).

One possible way to accomplish the mission of spreading the word about corpora is to train future language teachers in their pre-service training, as expressed by many scholars of the field, cf. Meunier (2002); O’Keefe & Farr (2003); Mauranen (2004); Römer (2006); Granath (2009); Boulton & Tyne (2014). Student (or trainee) teachers, as the very name suggests, are in the peculiar position of still being students while progressing to becoming teachers. Thus, not only do they have first-hand language learning experience but they can already see the pedagogical value of using corpora. This fact has been recognised by a growing number of teacher training programmes across the globe. They incorporated corpora in their curriculum either as part of pre-existing Pedagogical Grammar courses (Helt & Reppen, 2008; Heather & Helt, 2012; Samburskiy, 2014; Zareva, 2016), or in the form of introductory training on corpus applications in language education (Farr, 2008; Breyer, 2009; Leńko-Szymańska, 2014, 2015; Callies, 2016; Naismith, 2016).

At the same time, while some professors mentioned through personal communication that they sometimes use corpora to enhance their academic courses, a systematic and practical introduction to corpora is still to be done in the Hungarian teacher training context. For this reason – drawing on the results of the aforementioned studies –, with my thesis supervisor’s consent, I decided to bring a change to the status quo in Hungary. Two practical workshop sessions were organised on the uses of corpora in language teaching through pedagogical grammar-related tasks at a large public university in Budapest. The workshops were held within the framework of Pedagogical Grammar courses in the MA in English Language Teaching programme with 33 Hungarian trainee teachers of English.

Purpose of the study

The purpose of this pioneering and exploratory case study was to assess trainee teachers’ prior knowledge about corpora and corpus linguistics and, after the two workshop sessions, to evaluate the effectiveness of these workshops regarding the future, based on questionnaire data from trainees and their home assignments. These aims motivated the research questions of the study:

- How can trainee teachers’ prior knowledge about corpora and corpus linguistics be described?
- How effective were the workshops as perceived by trainee teachers?

The case

The two workshops represented the case of the study. They were offered in two separate, 90-minute-long sessions to two groups of trainee teachers, and taught by the author. The
two sessions were identical in content, namely both of them revolved around the following three major topics: corpus linguistics, language teaching and pedagogical grammar.

The design of the workshops followed Mukherjee’s (2004) proposed framework for a one-day workshop, and they consisted of three main, sequential thematic modules:

(1) teaching about corpora – a brief presentation of the basic concepts;
(2) exploiting corpora to teach language – various hands-on explorations in the BYU corpora\(^1\) with the help of five topics from prescriptive grammar (e.g., the majority is or are, according to me…) and five from descriptive grammar (e.g., help sb (to) verb, double negatives)\(^2\);
(3) teaching to exploit corpora – discussion about classroom applications of corpora, supported by some relevant books, such as O’Keeffe, McCarthy & Carter (2007), Tribble & Jones (1997) and Willis & Willis (1988).

Methods of data collection and analysis

As the research is a case study, one of the top priorities was to use multiple data sources. The data analysed in the study came from the following three major sources:

(1) a questionnaire before the workshop – to explore trainees’ level of knowledge about corpora and corpus linguistics prior to the instruction;
(2) participants’ home assignments – to assess trainees’ ability to make targeted searches in the corpus, interpret corpus data and draw pedagogically relevant conclusions;
(3) a questionnaire after the workshop – to get feedback from trainees on their perception of their own development after the workshop.

Both questionnaires included 5-point Likert scale items and open-ended questions, and their structure was based on those administered by Leńko-Szymańska (2015). In the case of each Likert scale item, after assigning each response option a numerical value (i.e., 5 to strongly agree, 4 to agree, …, 1 to strongly disagree), the mean and the standard deviation were calculated to specify the mean trend. With open-ended questions, the ocular scan method was used (Russell Bernard, 2000).

Home assignments as principal qualitative documents were analysed in an inductive way. All home assignments were eyeballed and labelled. At this point, several common themes started to emerge, which were used to identify the main trends.\(^3\)

During the process of data analysis, sources of data were triangulated in order to determine whether data from the questionnaires and those from the home assignments pointed to the same conclusions.

---

\(^1\) [http://corpus.byu.edu](http://corpus.byu.edu)

\(^2\) These task ideas were inspired by Professor Mark Davies’ teaching materials at Brigham Young University, publicly accessible via [http://davies-linguistics.byu.edu/ling485/](http://davies-linguistics.byu.edu/ling485/).

\(^3\) Here, as in Zareva’s (2016) study, a trend was defined as at least three responses with identical content.
Results

The findings of the study can teach us several invaluable lessons as regards future integration of corpora and corpus linguistics into teacher training programmes. In the following, results of the three data sources will be presented and discussed in turn.

1 Pre-workshop questionnaire

The pre-workshop questionnaire consisted of three statements (Q1 I have already heard about corpora and corpus linguistics, Q2 I have a fair understanding of corpora and corpus linguistics, Q3 I have already had a chance to use corpora). The means of responses to these statements showed a downward trend, which can be visualised as in Figure 1.

What does this graph tell us? Over 77% of the respondents had some prior knowledge about corpora, but only around half of them indicated that they knew corpora reasonably well. On the other hand, only 3 trainees, out of 33, signalled that they had had the chance to try corpora out, resulting in a mean of 2.23. Furthermore, none of them had exploited corpora previously for searching for any kind of grammatical structures.

![Graph showing downward trend of means from statements of the pre-workshop questionnaire.]

*Figure 1.* The downward trend of means from statements of the pre-workshop questionnaire.

2 Home assignments

Trainees’ home assignments showed a more mixed picture. Using the eyeballing and labelling technique, four general themes had been established, which are shown in Table 1.

<table>
<thead>
<tr>
<th>Trends</th>
<th>No of assignments</th>
</tr>
</thead>
<tbody>
<tr>
<td>incomplete assignments</td>
<td>7</td>
</tr>
<tr>
<td>diverse approaches (e.g., in the double negation task)</td>
<td>4</td>
</tr>
<tr>
<td>critical thinking and sophisticated solutions</td>
<td>3</td>
</tr>
<tr>
<td>decent job</td>
<td>3</td>
</tr>
</tbody>
</table>

*Table 1.* Trends emerging from the home assignments.
A group of trainees submitted incomplete home assignments. Even if they could obtain raw data from the corpus, they did not attempt to draw conclusions from them. What is to be learnt from this group of assignments? To help trainees overcome these initial struggles, instructors of future offerings should emphasise the goal-orientedness of corpus searches and offer trainees even more guidance throughout the process.

Nevertheless, there were a handful of trainees who solved their tasks in a scientific and professional way, providing a balanced opinion about grammatical topics. They exhibited not only deep knowledge about the particular topic but also advanced research skills, which corresponds to what was found by Farr (2008) and Callies (2016). As one trainee noted in connection with the according to me... vs in my opinion problem, "In my opinion is much more frequent than according to me. In COCA, there are very few instances [when] according to me is used. I would not enforce teaching according to me, as in everyday life the students will come across in my opinion."

3 Post-workshop questionnaire

Thanks to the workshops, trainees expressed that they had become more aware of available corpus resources and tools (M = 3.91) and that they had come to know how to use corpus resources fairly well (M = 3.41). However, some of the trainees are still not very confident to interpret corpus data on grammar (M = 3.05), which is congruent with the most common trend from the home assignments.

Furthermore, over three thirds of the trainees were keen on learning more about using corpus tools in education (M = 3.68), but some turned out to be still hesitant to exploit corpora in their own teaching (M = 2.36). This was due to the limitation of the study, as two one-off workshop sessions could not possibly show all aspects of using corpora in classrooms. Nevertheless, future sessions will be organised to accommodate this need as well.

Conclusions

The findings of the study have practical implications for any teacher training programme aiming at having corpus linguistics as its component. There are at least three possible paths that can be taken from here. One option is to fully integrate corpora into Pedagogical Grammar courses, as was done by Helt & Reppen (2008), Heather & Helt (2012), Samburskiy (2014) and Zareva (2016). An alternative to this is to introduce a separate Introduction to Corpora and Corpus Linguistics course to the teacher training programme (either obligatory or elective), which would be akin to the courses described in the studies by Farr (2008), Breyer (2009), Lęńko-Szymańska (2014, 2015), Callies (2016) and Naismith (2016). A third route would include a series of workshop sessions organised for trainees or even for practising teachers, an example of which was described in this study.

My vision, in line with that of the authors of previous research studies, is about corpus literate language teachers. Teachers who are able to make better decisions in their teaching as a result of their ability to manipulate corpora skilfully and provide more appropriate answers to their students based on authentic language data. To achieve this, we need to make the first step forward on either of the above paths.
References


Meunier, F. (2002). The pedagogical value of native and learner corpora in EFL grammar teaching. In S. Granger, J. Hung & S. Petch-Tyson (Eds.), Computer learner corpora, second language acquisition and foreign language teaching (pp. 119–141). Amsterdam, NL: John Benjamins.


Premodifying female and male adjectives in journalistic texts.  
A gender corpus analysis in Czech  
Adrian Jan Zasina (The Czech National Corpus, Charles University, Czech Republic)

Currently, there are used a range of methods in the discourse analysis of gender (Baker 2014: 4) but most of these studies have been conducted on English. Our analysis engaging Czech data is one of the first in the field to analyse gender in Czech. In this research, we would like to describe the corpus approach to the analysis of gender in Czech and present similarities and differences between men and women as they appear in journalistic texts.

In the Czech environment, there have been previous qualitative gender studies (Hofmannová 2004) but a corpus approach to analysing gender remains unpopular. Therefore, our research was inspired by some previous discursive gender analysis of collocations conducted by Pearce (2008), Caldas-Coulthard & Moon (2010), Baker (2010), Macalister (2011), Taylor (2013), Moon (2014).

Research material and tools

For purpose of this study, we used a representative selection from the contemporary written Czech SYN2015 corpus (Křen et al. 2015; Křen et al. 2016; Cvrček, Čermáková & Křen 2016) which is divided into three text types: fiction (FIC), non-fiction, and newspapers and magazines (NMG). In our investigation, we took into account NMG, using a research sample which consists of 39,744,419 tokens.

For data analysis, we used two different interfaces: KonText (Machálek & Křen 2013) and a paradigmatic query interface (Cvrček 2017). KonText is an open-source application developed within the framework of the Institute of Czech National Corpus (ICNC, http://ucnk.ff.cuni.cz). The interface for paradigmatic query is not formally available as it is still in development at ICNC. This kind of search tool made it possible to classify adjectives into groups that collocate exclusively, almost exclusively or predominantly with the lexemes muž and žena.

Analysis

In our analysis, we dealt with the lexemes žena/muž in the journalistic part (NMG) of the SYN2015 corpus. We were looking for premodifying adjective collocates. In the next step, the adjectives we found were divided into groups according to exclusive, almost exclusive, or predominant appearance with one or another lexeme. Based upon this, we established four groups each sorted out into masculine (collocating with lexeme muž) and feminine (collocating with lexeme žena) as well and one group of adjectives common for both sexes.

Based on these four groups, we created ten semantic categories into which we sorted out the adjectives. These ten categories were labelled as follows: age, strength and supernatural power, appearance and attractiveness, character, psychological state and adjectives evoking positive/negative emotion, maternity, nationality/ethnicity, action, marital status, sexual orientation and others.
Our analysis has shown that a stereotypical view of men and women is still present in journalistic texts. We proved that certain adjectives are used only with one of the examined lexemes (e.g. *klíčový* 'key' only with *muž* and *practical* 'practical' only with *žena*). Our analysis shows that in the case of masculine adjectives, men are often described by age, strength and supernatural power. In the case of feminine adjectives, they are often related to motherhood and nationality/ethnicity. Adjectives evoking negative and positive emotions, appearance and attractiveness are common for both.

References

Lexis and Tradition: Variation in the vocabulary of Buddhist Literature
Ligeia Lugli (King’s College London, UK)

This poster reports on the initial phase of a British Academy-funded project on lexico-semantic variation in Buddhist Sanskrit. It outlines the overall scope and aims of the project, which addresses three sets of questions:

(1) Is there a discernible pattern in the distribution of near-synonyms across Buddhist texts and traditions? If there is, to what extent may this point to different communities of practice, and thus provide a clue to self-identity within Buddhist traditions?

(2) To what extent inter-sectarian dialogue is likely to have affected lexical choices and led to a degree of terminological harmonisation across different schools?

(3) To what extent doctrinal and conceptual evolution drove lexico-semantic innovation? In particular, to what extent might the constraints of orthodoxy have induced semantic change in the religio-philosophical vocabulary by promoting the re-contextualisation of traditional wording patterns in new doctrinal frameworks?

To answer these questions a diachronic corpus of Buddhist Sanskrit texts is being compiled. Within this corpus, lexico-semantic variation will be analysed using the methods for diachronic onomasiology described by Geeraerts (especially 2002 and 2012). The research will focus on two thematically related conceptual fields, NAMING and CONCEPTUALISING. To track semantic variation within these fields, the corpus is being semantically tagged using a word-sense annotation system derived from the taxonomy of the Historical Thesaurus of English and adapted to Sanskrit (Lugli 2015).

The poster also reports on the progress made in the development of a diachronic corpus of Buddhist Sanskrit. It describes the challenges that Buddhist Sanskrit materials pose for corpus linguistics and how these challenges have been addressed within this project, focussing especially on the difficulties related to the automated tokenisations and lemmatisation of Buddhist Sanskrit, the problems of corpus design that arise from the absence of a clear relative chronology for Indic Buddhist texts, and the task of adapting an existing semantic tagset to cultural-specific concepts of South Asian Buddhism.

Finally, the poster presents preliminary findings on a Buddhist Sanskrit keyword, *saṃjñā*, and summarises how corpus data about this word problematise current assumptions and translation practices in the field of Buddhist Studies.
References


This paper extends the concept of local grammar (Hunston & Sinclair 2000) to speech act studies (Austin 1962; Searle 1969). The reason for doing this is that, while Butler (2004: 158) has noted that “rather than a single general grammar, we might end up with a set of local grammars for particular areas defined by their communicative functions in the discourse”, very few studies have investigated empirically the possibility of doing so. Since speech acts are generalisations of communicative functions, it is arguable that an investigation into local grammars of speech acts would help us to explore the feasibility of developing a set of local grammars to account for language used in social interactions. This study therefore presents a preliminary investigation into local grammars of speech acts, focusing specifically on apologising. The primary aim is to develop a local grammar of apology and, based on which, to further explore the possibility of using local grammars to account adequately for speech acts; and the ultimate aim is to explore the general applicability of local grammars in linguistic description and explanation.

Simply put, local grammar is an alternative approach, as opposed to general or traditional grammars, to the description and theorising of language in use. The defining features of local grammar include, first, each local grammar deals with one meaning or function only, and second, it involves mapping functional elements on to formal/pattern elements, thus facilitating the establishment of the connection between form and function in interactive contexts. Third, local grammar takes into account the functions language fulfils in social contexts and analyses each discourse unit in terms that are related directly to its discourse function; and as such, local grammar is in essence a functional description of language in use. To date, local grammars have been applied to the study of, for example, definition (Barnbrook 2002), evaluation (Bednarek 2008; Hunston & Sinclair 2000; Su 2015), disclaimers (Cheng & Ching 2016), and request (Su 2017). Additionally, Warren and Leung (2016) also extends local grammars to describe patterns of co-selection found in collocational frameworks. In general, what these studies have shown is that local grammars can provide a more systematic and comprehensive description of one particular meaning or function, which indicates the significance of the construction of local grammars.

The other framework this study draws on is speech act theory which generally means that in saying something we are also doing something. The particular type of speech acts being focused on is that of apologising; this is because: first, apology is a ritual work that is important for maintaining interpersonal rapport as well as for restoring social equilibrium and harmony; and second, many studies have shown that they are realised by more or less fixed, recurring patterns, which makes it relatively easier to maximally identify instances of apologies in a corpus.

The corpus used in this study is compiled of transcripts of the first seven seasons of the sitcom The Big Bang Theory. Transcripts of sitcoms are used in the present study partly because it is very difficult to get sufficient amount of data by
recording conversations in real contexts, and more importantly, because conversations between the characters in sitcoms have high similarities with our daily or casual conversation in which the speech act of apologising is frequently performed (cf. Quaglio 2009). The corpus of The Big Bang Theory compiled (henceforward CBBT) comprises 159 texts and has 485,602 tokens. The corpus was uploaded to Sketch Engine (Kilgarriff et al 2004) for further processing.

This study focuses on apology expressions containing conventionalised forms of apologies. The rationale behind this is that “[a]pologies are generally made up of a small repertoire of relatively fixed expressions representing verbs (apologize, excuse, pardon), adjectives (sorry, afraid) and nouns (pardon) and their expansions, modifications” (Aijmer 1996: 84). Drawing on insights from previous investigation into apology (Aijmer 1996; Blum-Kulka et al. 1989; Deutschmann 2003; Jucker & Taavitsainen 2008), this study uses the following key terms to search and extract instances of apology in the CBBT; the forms and their quantitative information are given in Table 1.

The analyses are divided into six sets, according to their similarities and degree of complexity. Table 2 presents an overview of the local grammar patterns of apology identified, with an example given for each pattern. It can be seen that strategies for apologies can be divided into two broad categories, i.e. making apologies and seeking forgiveness. Furthermore, the quantitative information suggests that the former is the typical way of apologising.

Based on the analyses, we can summarise the set of functional labels that are needed for a local grammar analysis of apology (Table 3). It should be admitted, though, that it is not clear whether these labels are comprehensive and sufficient enough for analysing all kinds of apology expressions, as this study has only investigated apology expressions containing conventionalised forms. Nevertheless, since many studies have shown that realisations of apologies are highly conventionalised (Aijmer 1996; Deutschmann 2003), it is reasonably confident that the key terms listed in Table 1 would enable us to identify maximally instances of apologies. This further indicates that these labels would be able to account for most apology expressions in real contexts, because they are not thought-up, but proposed on the basis of analysing instances containing those key terms.

To conclude, the study proposed that a local grammar approach would be useful to account for speech acts. It has reported an investigation into the local grammar of apology, which, we believe, has amassed sufficient evidence to show the possibility and feasibility of developing a set of local grammars to account more adequately for speech acts. Two characteristic advantages of local grammars are particularly worth recapitulating. First, local grammars use context-specific functional element to analyse each corresponding formal element; the resulting description is therefore transparent and a real functional account of language in use (Hunston & Sinclair 2000). Second, local grammars are simpler, compared with general grammars, in that each local grammar deals with only one meaning or discourse function. In the case of speech acts, each local grammar accounts for one particular speech act. Although this indicates the loss of generalisability of the description, this is compensated for by the gains of cumulative coverage achieved by a set of local grammars. An extension of this argument is that local grammars are indeed of general applicability in linguistic description and explanation.
<table>
<thead>
<tr>
<th>Item</th>
<th>Realisation</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>SORRY</td>
<td>sorry</td>
<td>197</td>
</tr>
<tr>
<td></td>
<td><em>I/we v-link</em> sorry</td>
<td>295</td>
</tr>
<tr>
<td></td>
<td><em>I/we v-link</em> sorry for/about/that/to-inf.</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>sorry about/that/to-inf.</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td><em>I/we v-link</em> intensifier sorry</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td><em>I/we v-link</em> intensifier sorry for/about/that</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td><strong>Subtotal:</strong></td>
<td>679</td>
</tr>
<tr>
<td>EXCUSE</td>
<td>excuse me</td>
<td>194</td>
</tr>
<tr>
<td></td>
<td>will/can you excuse me/us/sb</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>excuse me for</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td><strong>Subtotal:</strong></td>
<td>212</td>
</tr>
<tr>
<td>APOLOGIZE/SE</td>
<td><em>I/we (v-link)</em> apologize/se</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td><em>I/we (v-link)</em> apologize/se for</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td><em>I/we (v-link)</em> apologize/se to NP</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td><strong>Subtotal:</strong></td>
<td>31</td>
</tr>
<tr>
<td>APOLOGY</td>
<td>my apologies</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>My apologies for</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>My apologies to NP</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td><strong>Subtotal:</strong></td>
<td>14</td>
</tr>
<tr>
<td>FORGIVE</td>
<td><em>(please)</em> forgive me</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>forgive me for</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>forgive my (language/crude penmanship)</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td><strong>Subtotal:</strong></td>
<td>14</td>
</tr>
<tr>
<td>REGRET</td>
<td><em>I regret</em> something or not doing something</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td><strong>Subtotal:</strong></td>
<td>8</td>
</tr>
<tr>
<td>AFRAID</td>
<td><em>I'm afraid</em> ...</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td><strong>Subtotal:</strong></td>
<td>7</td>
</tr>
<tr>
<td>PARDON</td>
<td>pardon me</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td><em>(I) beg</em> your pardon</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td><strong>Subtotal:</strong></td>
<td>5</td>
</tr>
<tr>
<td></td>
<td><strong>TOTAL:</strong></td>
<td><strong>970</strong></td>
</tr>
</tbody>
</table>
**Table 2.** An overview of local grammar patterns of apology

<table>
<thead>
<tr>
<th>Analyses</th>
<th>Patterns</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Set 1</strong></td>
<td>Apologising</td>
<td>e.g. <em>sorry</em></td>
</tr>
<tr>
<td></td>
<td><strong>Apologising + Specification</strong></td>
<td>e.g. <em>sorry for being late</em></td>
</tr>
<tr>
<td></td>
<td><strong>Apologising + Apologisee</strong></td>
<td>e.g. <em>my apologies to you all</em></td>
</tr>
<tr>
<td></td>
<td><strong>Subtotal:</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Set 2</strong></td>
<td>Apologiser + Hinge + Apologising</td>
<td>e.g. <em>we’re sorry</em></td>
</tr>
<tr>
<td></td>
<td>Apologiser + Hinge + Intensifier + Apologising</td>
<td>e.g. <em>I’m really sorry</em></td>
</tr>
<tr>
<td></td>
<td><strong>Subtotal:</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Set 3</strong></td>
<td>Apologiser + Hinge + Apologising + Specification</td>
<td>e.g. <em>I’m afraid we can’t authorize that</em></td>
</tr>
<tr>
<td></td>
<td>Apologiser + Hinge + Intensifier + Apologising + Specification</td>
<td>e.g. <em>I’m truly sorry for what happened</em></td>
</tr>
<tr>
<td></td>
<td><strong>Subtotal:</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Set 4</strong></td>
<td>Apologiser + Apologising</td>
<td>e.g. <em>I apologize</em></td>
</tr>
<tr>
<td></td>
<td>Apologiser + Apologising + Specification</td>
<td>e.g. <em>I apologize for my earlier outburst</em></td>
</tr>
<tr>
<td></td>
<td>Apologiser + Intensifier + Apologising + Specification</td>
<td>e.g. <em>I do regret not following up …</em></td>
</tr>
<tr>
<td></td>
<td><strong>Subtotal:</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Set 5</strong></td>
<td>Apologiser + Hinge + Apologising + Apologisee</td>
<td>e.g. <em>I wanted to apologize to the two of you</em></td>
</tr>
<tr>
<td></td>
<td><strong>Subtotal:</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Set 6</strong></td>
<td>Apologiser + Hinge + Forgiveness-seeking + Apologiser</td>
<td>e.g. <em>(hope) you can forgive me</em></td>
</tr>
<tr>
<td></td>
<td>Forgiveness-seeking</td>
<td>e.g. <em>forgive me</em></td>
</tr>
<tr>
<td></td>
<td>Forgiveness-seeking + Apologiser + Specification</td>
<td>e.g. <em>excuse me for being so bold</em></td>
</tr>
<tr>
<td></td>
<td><strong>Subtotal:</strong></td>
<td></td>
</tr>
<tr>
<td><strong>TOTAL:</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 3. Functional labels for analysing apology

<table>
<thead>
<tr>
<th>Element</th>
<th>Explanation</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apologiser</td>
<td>The one who apologises</td>
<td>I am sorry.</td>
</tr>
<tr>
<td>Apologising</td>
<td>The elements that realise apologies</td>
<td>I apologize.</td>
</tr>
<tr>
<td>Forgiveness-seeking</td>
<td>The action of seeking forgiveness</td>
<td>Please forgive me</td>
</tr>
<tr>
<td>Apologisee</td>
<td>To whom the apology is made to</td>
<td>Just apologize to him.</td>
</tr>
<tr>
<td>Intensifier</td>
<td>The elements that upgrade the degree of regret</td>
<td>I am so sorry.</td>
</tr>
<tr>
<td>Specification</td>
<td>The elements that specify the reason for an apology</td>
<td>I am sorry for what I said.</td>
</tr>
<tr>
<td>Hinge</td>
<td>The elements that link different functional elements</td>
<td>I am really sorry about this.</td>
</tr>
</tbody>
</table>

References

Animacy hierarchy within inanimate nouns: English corpus evidence from a prototypical perspective

Jie Ji (China Foreign Affairs University, China) and Maocheng Ling (Beijing Foreign Studies University, China)

Animacy has usually been regarded by linguists as a gradient hierarchy rather than a simple biological dichotomy. The animacy of an entity depends not on its biological attributes, but on “the speaker’s identification or empathy” with this entity (Kuno and Kaburaki, 1977: 628), or “to what extent speaker treat this entity as if it was animate” (Rosenbach, 2008: 154). Animacy hierarchy can be seen as the gradient assessment towards all entities from the speaker’s point of view, and it proves to be widely existent in human languages and highly influential to many grammatical phenomena. Typological studies find that animacy hierarchy has exerted considerable constraints on phenomena like case, number, agreement, voice and syntactic order (Aissen, 2003; Comrie, 1989; Corbett, 2000; Jelinek and Demers, 1983; Oshima, 2007), and these constraints prevail in human languages as either grammatical rules or linguistic preferences (Bianchi, 2006; Bresnan et al., 2001; Rosenbach, 2008).

Drawn from those linguistic manifestations or cognitive assessments, several animacy hierarchies have been proposed (Corbett, 2000; Dixon, 1979; Langacker, 1991; Silverstein, 1976; Yamamoto, 1999). These hierarchies are fine-grained only within the animate entities. Person pronouns, proper nouns, human-referring nouns and animals are granted different ranks along the scale, leaving the inanimate entities “an undifferentiated class” with only few “arbitrary” distinctions (Comrie, 1989: 197). Does this mean that all inanimate entities are equally inanimate without any hierarchic nature? Or that their hierarchy is simply not worth exploring for its insignificant impacts on human languages? These are the two questions the present research is going to investigate.

According to the cognitive approach (Dahl, 2008; Langacker, 1991; Yamamoto, 1999), animacy hierarchy reflects human’s egocentric or prototypical assessment of entities. Animacy of an entity depends on how much resemblance it bears to us, or more specifically, how many human prototype features it possesses. The animacy of inanimate nouns could be gradient if they are assigned different amounts of human features by us. Human is agentive, definite, individual, kinetic, cognitive and social (Aissen, 2003: 437; Dahl, 2008; Dixon, 1979: 85; Gelman et al., 1995; Johnson et al., 1998; Woodward, 1998). However, these simple lists of prototype features are inadequate to predict a credible hierarchy, as the amount and strength of each category’s features are difficult to measure in this way. We need a feasible method to externalize and quantify these prototype features.

When we say humans are physically kinetic or psychologically cognitive, sentient, communicative and social, we mean they can walk, think, love, exchange ideas and socialize with each other. In other words, humans are “teleologically capable of generating these actions” (Folli and Harley, 2008: 192), that is, humans have the inherent attributes to participate in the actions described by these predicative verbs. In this sense, human attributes or the prototype features can be externalized linguistically as human subjects’ predicative verbs. Whenever a human subject occurs with a predicative verb, we can add that verb to the feature list. Then the amount of prototype features can be easily quantified as the total occurrences of these verbs.

This method can calculate not only the amount, but also the strength of prototype features. For example, humans can think and change, thus the quality to “think” and
“change” can both be regarded as human attributes or prototype features. However, the two features have different weights or “cue validities” in their contribution to animacy. Just as “gills” is more valid than “oviparity” when labelling the fish category (Lakoff, 1987: 53), the ability to “think” is much more valid than the ability to “change” when identifying human category. The validity of each cue depends on how frequently “it is associated with members of that category” (Evans and Green, 2006: 261). Thus, the weight of each prototype feature can be calculated by how often its externalized predicate co-occurs with human subjects. More weighting should be given to those more valid features when calculating the amount and strength of each noun category’s prototype features.

By using predicative verbs as linguistic externalization of prototype features, this research intends to investigate the animacy hierarchy within inanimate nouns through human language. English is used as the experimental language in this study. The corpus data come from the written part of British National Corpus (BNC). WordNet is used as a noun database in this study. It classifies all nouns into 25 groups, and lists each sense of each noun with group label and corpus frequency. This is useful for our noun grouping procedure. Stanford Parser is used to parse grammatical relations for corpus data. It can provide dependencies such as “nsubj”, which labels the syntactic subject and predicate in a clause. We can use this dependency to calculate how often each predicative verb co-occurs with human subjects (hence the weight of each feature), or how frequently and strongly each inanimate noun category co-occurs with these verbs (hence the animacy of each category). We also develop an animacy calculating software that can search through the parsed corpus and automatically calculate the occurrence, weight and animacy, then finally provide a Microsoft Excel output according to our design.

The results attest to the very existence of the animacy hierarchy of five inanimate categories. Collective nouns, spatial and temporal nouns, concrete nouns, psychological nouns, and other abstract nouns rank in descending sequence along the hierarchy. This hierarchy is determined directly by each category’s feature distribution in different weight intervals, and the semantic distribution helps to understand the underlying cognitive mechanism of this hierarchy. Further syntactic exploration show that this hierarchy can impact on word order just as those existing hierarchies do.

Reference


This study looks at how predicate adjectives and nouns are used with the pronouns ‘we’ and ‘they’ in discourse to produce and maintain social categories relative to the speaker. Pronouns have been found to be relevant in the establishment of social distance between participants depending on the pronoun being used (Fitzsimons & Kay, 2003). For example, a listener might perceive the participants to be closer in a story if the speaker uses the pronoun ‘we’ rather than the participants’ individual names. Similarly, the first person plural pronoun ‘we’ and third person plural pronoun ‘they’ can have similar effects with regards to the inclusion or exclusion of the speaker. If the first person plural pronoun ‘we’ is used the speaker aligns themselves with the group in question whereas the use of ‘they’ distances the speaker from the group. This process along with the choice of predicate adjectives and predicate nouns allows speakers to produce different social categories through their speech. Predicate adjectives and nouns are those that follow the copula in a simple declarative sentence (Greenbaum & Quirk, 1993) such as in the sentence ‘we are women’ where the noun ‘women’ follows the copula ‘are’. These predicate adjectives and predicate nouns were analyzed specifically, because it has been found that these copular constructions are highly evaluative in nature providing a lot of the speaker’s opinions about the subject matter (Gianluca & Stanislaw, 2014). Thus, analyzing this process will help us to better understand how social categories are produced and maintained in a social setting.

For this study, the data was collected from the Contemporary Corpus of American English (COCA). Because this research was looking for category construction in discourse, only oral data was used. All of the adjectives that appeared after ‘we are’ and ‘they are’ were recorded and coded for their frequency and context. The predicate nouns were recorded by searching for nouns following ‘we are’ and ‘they are’ but a word was permitted to separate the noun from the copula to account for articles. These, too, were coded for frequency and which pronoun was used. While it is possible to have a noun phrase or adjective phrase that is longer than two words, for this study simple noun and adjective phrases were used. Upon the completion of the data collection, the adjectives were further coded for the type of adjective according to Dixon’s categories: human propensity, size, physical property, speed, value, etc. (Dixon, 1977). In a sentence like, ‘we are concerned’ for instance, the adjective ‘concerned’ would be categorized as human propensity because it is describing an emotional or mental state. The nouns were categorized into group words, individual words, and generalizations. Group words include any words that refer to a group like ‘team’ while words like ‘leader’ or ‘individual’ were labeled as individual. The category labeled generalization was for any noun that referred to a social category. This was chosen specifically because generalizations have been found to be highly evaluative (Scheibman, 2007). For example, looking at the statement, ‘we are women’ again, the predicate noun is referring to a specific social category, women.
A total of 7,228 tokens were gathered with 3,251 tokens in constructions with ‘we’ and 3,977 appearing with the pronoun ‘they’. Adjectives had 5,268 (73%) occurrences while nouns only had 1,960 (27%) occurrences. However, the nouns had a wider variety with 160 different nouns being used and adjectives only having 121. Among the adjectives, the three most common types of adjectives used were human propensity, physical property, and value. The percentages for these three can be seen in the following chart which indicates that human propensity was the most common with the first person pronoun at 57% and physical property was the most common for the third person pronoun with 47%.

For the nouns, the tokens were divided up into ‘group’ words, ‘individual’ words, and words that referenced social groups. For these three categories, group words were used significantly more with the first person plural pronoun with 62% compared to 28% for ‘they’. However, the inverse was true for words emphasizing the individual which occurred more frequently with the third person plural pronoun, 63% compared to 32% for ‘we’.
This data illustrates the process by which social categories are administered to particularly with relation to the speaker. As individuals speak, they use the two pronouns, ‘we’ and ‘they,’ to either align or distance themselves from the material at hand. When they do include themselves in the group being discussed there is an emphasis on the group which can be seen in the nouns that are used for these constructions. In contrast, when the speaker does not attach themselves to the subject matter and uses the pronoun ‘they,’ there is a preference to use nouns emphasizing the individual. Once these groups are established and the speaker’s relation to them is addressed, further description occurs with the use of predicate adjectives. With the predicate adjectives, there is a higher tendency of inclusive groups to describe their feelings with adjectives in the human propensity category while the exclusive categories rely more heavily on physical properties to define them. The actual description of the groups provided by the adjectives appears to be of greater import than the initial labeling based on the significantly higher occurrences of predicate adjectives than predicate nouns.

By understanding this process of category production and maintenance, we are better able to understand how individuals use language as a tool to navigate through society and establish their own identity. When these social categories are established, the individual’s identity can be seen to emerge, as well, through its relation to these categories. In this way, both social categories and social identities are produced and maintained through our discourse.
References


One of the purposes of disability studies, in addition to advocating for the rights of disabled individuals, is to provide a thoughtful perspective on works of art and literature with the goal of focusing attention on their portrayals of disabled people, which can both reflect and shape societal attitudes. Attending to these portrayals, then, becomes a means of examining—and in some cases opposing—such attitudes. (See Longmore, 2003, for a useful introduction to disability studies.)

Often, such examinations are focused on agency; i.e., do the disabled characters wield control over their own lives, or are they controlled by others? A traditional analysis of this sort can be done by selecting scenes or passages from the text and using the tools of literary criticism to explicate them. This approach can obviously lead to controversy; analysts can be accused of ‘cherry picking’ particularly egregious scenes or avoiding portions of the text that appear to contradict their analysis. Here, we employ techniques drawn from corpus linguistics and ranging over an entire text, to establish a firmer basis for claims about a character’s agency.

The novel All the Light We Cannot See, by Anthony Doerr (2014), is a particularly cogent example. It has been widely praised, winning the Pulitzer Prize in 2015; in particular, the presence of a blind character, Marie-Laure, has been viewed positively (Faerber, 2015). At the same time, negative reviews of the book cite it as a classic example of destructive stereotypes about blindness (Wells-Jensen, 2016).

Using traditional methods, advocates for the novel could point out passages where the blind main character volunteers for a task to assist the French Resistance and to a scene at the end of the novel where she is portrayed as receiving a graduate degree. On the other hand, critics who judge the book as ableist might point out a quasi-incestuous scene where the father washes Marie-Laure’s hair and numerous occasions where other characters assist her in putting on her shoes or “button her into her coat”. Both sides eventually may come to a standstill as isolated pieces of ‘evidence’ are offered, disputed, reanalyzed, and eventually rejected.

Usefully, constructs such as ‘agency’ have lexical, semantic, and syntactic correlates which can be examined throughout the entirety of the book, avoiding accusations of partisan data mining. These correlates can be located, counted, and discussed with some measure of objectivity. This paper lays out a method for engaging a text in this way.

Thus, the question asked is not “Does this book, as represented by these selected scenes, promote destructive stereotypes of disabled people?”, but rather “Do the words, syntactic frames, and illocutionary forces used consistently
throughout the book to portray this character differ in measurable ways from those representing non-disabled characters appearing in the same text?"

In addition to some simple in-house Perl scripts for sorting and counting words, we used the Stanford CoreNLP 3.7.0 toolkit (Manning et al., 2014) as a starting point. CoreNLP produces parses and pronominal coreference chains, crucial for determining who is involved in events and in what capacity. To assist in the determination of lexical items to be examined, we elected to use the Harvard General Inquirer (General Inquirer, n.d.), as it contains a plethora of categories relevant to the task, including rankings of words as ‘strong’, ‘weak’, ‘passive’, and ‘active’. At this point, many detailed investigations became possible:

• We collected conversational turns, associated each with the character responsible for the utterance, and labeled each with its simple illocutionary force. This enabled us to determine, e.g., which characters issued commands or asked for information. Roughly 60% of Marie-Laure’s utterances are requests for information, as compared to an average of 28% for other characters. Conversely, only 5% of her utterances are commands, while utterances for other characters range from 10% to 20% commands.

• We also labeled characters’ utterances by addressee. Although it is very common for Marie-Laure to ask questions, only 10% of requests for information in the book are addressed to her. Furthermore, half of these are requests for information about her state of health, while the questions addressed to other characters focus on information they possess, events they have witnessed, or the degree to which they agree or disagree with the topic under discussion.

• Using Inquirer words categorized as ‘strong’ and ‘active’ which occur frequently in the novel, it is possible to locate their subjects, showing the degree to which the blind character is disproportionately the subject of verbs such as to feel and to beg. She does not appear as the subject for ‘stronger’ verbs such as to explain or to attack, and although there are 18 tokens of the verb to fight in the novel, she is the subject of only one of these—its object being temptation.

In terms of several linguistic and interactional variables, then, our analysis indicates that the blind character, Marie-Laure, is lacking in agency as compared to other characters in the same novel.

Of course, other variables are at play here; Marie-Laure’s age, her gender, her relatively high socio-economic status, and the fact that she is an only child and apparently without friends, all are aspects of her character which co-occur with her blindness. While acknowledging these things, one cannot then ignore the fact that her lack of agency reflects a persistent negative stereotype of blindness. It is remarkable how skillfully and consistently Doerr has—either consciously or unconsciously—crafted the character, forming a reliable lexical and syntactic foundation which supports this effect. If the rendering was consciously done, it is a remarkable example of attention to low level detail. If it was unconsciously done, and the intent was to create a strong character who defies conventional stereotypes of disability, the linguistic pattern of passivity is perhaps even more remarkable.
It is not our intention to create a new kind of imprimatur, a test by which a piece of writing can be either approved as ‘inclusive’ or dismissed as ‘ableist’. What we offer here is a set of reproducible, objective observations and techniques which can be used by both readers and authors to examine the substrate of any piece of writing. What they then choose to do with that knowledge is outside the scope of this paper.

References

Doerr, A. (2014). *All the Light We Cannot See*. Scribner.
Translation-oriented annotation of a multimedia parallel corpus of subtitles
Patricia Sotelo-Dios (University of Vigo, Spain)

This poster presents an ongoing research project that involves annotating a multimedia parallel corpus of subtitles with translation-related information. The Veiga corpus of English-Galician film subtitling currently covers 35 audiovisual products (ca. 450,000 tokens) and can be accessed freely at http://sli.uvigo.gal/CLUVI/vmm_en.html. The corpus search application allows for complex as well as parallel searches. It shows the bilingual equivalences of the searched terms in context and enables users to stream the video clips where the bilingual pair appears, thus giving them access to the (co-)text in its original, multi-semiotic form. The Veiga already contains two levels of annotation: on the one hand, the omission, addition and reordering of translation units; and, on the other hand, the in-cue and out-cue times and line breaks in the subtitles. All of the above mentioned aspects are tagged according to the XML CLUVI specification for parallel corpora (Guinovart & Sacau, 2004), and a new set of tags is now being defined to annotate certain elements that are particularly relevant to the practice of subtitling. The aim of the project is to enrich the corpus with translation-related data so that users can easily search for potentially problematic issues and observe the specific techniques used by the translator to render them in the subtitles.

A first pilot experiment is now being carried out with one of the films. In particular, the items currently being considered for annotation are the following: linguistic variation (dialect, slang, taboo language), idioms, culture-specific references, named entities, and orality features such as paralinguistic elements, discourse markers, interjections, vocatives, false starts or repetitions. Although a number of studies have already empirically investigated the translation of some of these issues in a variety of parallel subtitles corpora, with a special focus on the transfer of humour, cultural references, taboo language and linguistic variation, none of these corpora have yet been tagged with this type of information and made available to the public. In this respect, the Veiga corpus, with this additional layer of metadata, intends to make a small contribution to the annotation of parallel corpora for descriptive translation studies by providing a freely searchable database that could be used for translation teaching and research.

References

A corpus-based investigation of the lexical bundle use by accomplished and novice mathematics writers
Abdullah Alasmary (the Department of English Language and Translation, King Saud University, Saudi Arabia)

While there has been a growing body of research into lexical bundle use in a wide range of academic disciplines, little attention is given to domains of highly theoretical and abstract nature. Drawing on a 2-million-word corpus of doctoral dissertations, textbook chapters and peer-reviewed journal articles, this research is an attempt to fill this gap, comparing the use of recurrent bundles by published authors and graduate students in mathematics. A total of 291 four-word bundles that recur 25 times per million and appear in at least 15% of the texts were retrieved and their structural and functional attributes examined. Results show that graduate students tend to use fewer and less varied lexical bundles than textbook authors and article writers, an outcome that reinforces a pattern in the studies comparing accomplished authors and novice students in various academic disciplines. Structural analysis of the lexical bundles demonstrates that while the student writers rely more on phrasal patterns, the expert writers prefer bundles comprising clausal constructions. The analysis of the functional attributes of mathematical bundles reveals no difference between groups, with all mathematician writers relying more on research- and text-oriented bundles, but showing little interest in using participant oriented ones.

1. Introduction

Disciplinary writing constitutes a great proportion of semantically transparent and grammatically regular multiword units, which recur frequently, performing identifiable discourse functions and displaying analyzable structural patterns (Adel & Erman, 2012; Chen & Baker, 2010; Biber, Conrad, & Cortes, 2004; Cortes, 2006, 2013; Hyland, 2008a, 2008b; Pan, Reppen, & Biber, 2016). Several terms are used to refer to such units, including academic clusters (Hyland, 2008a), phrasal expressions (Martinez & Schmitt, 2012), lexical bundles (Biber et al., 2004; Pan et al., 2016), academic formulas (Simpson-Vlach & Ellis, 2010), multi-word constructions (Liu, 2012) and recurrent word combinations (Adel & Erman, 2012). While these terms can be placed under the umbrella of formulaic language (Wood, 2015; Wray, 2002), they are used here to refer to a special set of preassembled words whose meaning can be inferred from the meaning of their constituent parts and their identification in natural discourse involves applying some frequency and dispersion measures (Cortes, 2004).

While there is a great body of research investigating the use of lexical bundles in a variety of academic disciplines, mathematical discourse appears to be a notable exception. This may be due to the multilayered nature of the mathematical register which requires attention to other nonlinguistic means of
framing and advancing mathematical claims such as the symbolic and graphic representations accompanying a text (Huang & Normandia, 2007; O’Halloran, 2005; Schleppegrell, 2007). Another reason is the assumption that mathematics is the least language-dependent discipline, enough reason to impel researchers to examine bundles in rather supposedly language-rich contexts. Yet a third factor relates to the complex network of grammatical patterns characteristic of math discourse such as the use of highly technical lexis, patterns with special meaning, and discipline-specific rhetorical devices for describing underlying mathematical relationships, all of which seem to flummox non-specialists. Despite this obvious negligence, studies carried out in the field of mathematics education and English for Academic Purposes (EAP) concur that language plays an important role in how mathematical knowledge is constructed, disseminated and interpreted by the disciplinary members. As Graves, Moghaddasi and Hashim (2014) put it, “writing in mathematics is not different from other disciplines were a sophisticated awareness is required of communicative acts typical within that discourse community” (p.8). It is, therefore, plausible to employ linguistic tools to study corpus-derived patterns that contribute to knowledge making in a range of mathematics registers.

2. Corpora

Given that this study is comparative in nature, it was decided to create three parallel corpora comprising texts representing three academic registers: textbooks, peer-reviewed journal articles and PhD dissertations. The first corpus, named as TXT-BKs, includes chapters taken from textbooks aimed for graduate students in the field of mathematics. In order to guard against idiosyncratic use featuring some authors or texts, a decision was taken to include only one chapter from each textbook, with the final texts amounting to forty-three chapters. Each chapter is given a distinct code so as to trace the source of lexical bundles and to facilitate the process of the subsequent concordance checks.

The second corpus incorporates thirty-four peer-reviewed articles that appeared in the period between 2013 and 2015. Labelled as JOL-ARs, this corpus represents writings produced by notable authors and are published in leading, high-impact journals. The selection process of articles in this corpus was guided by three criteria. First, all journals from which articles are drawn must have a high impact factor and are indexed by the ISI Web of Knowledge. Second, only one article per issue and per author is included, so as to guard against issue or writer idiosyncrasies. Third, the content of each selected journal article must be in an electronic format so as to allow for its synthesis and analysis using automated corpus tools. Each article is cleared from appendices, references and acknowledgements.

The third corpus (DISS-ONs) is a collection of twenty PhD dissertations written by graduate students (10 males and 10 females). First language is highlighted in several studies as the cause of errors in the use of lexical bundles, so it was decided to control this variable by selecting the nonnative writings
produced by speakers whose first language is Arabic and for whom English is a foreign language. All the dissertations analyzed in this study are presented as texts upon which the Doctorate Degree was awarded to candidates who were pursuing their studies at English-only institutions and in countries where English is assigned the native language of the population. Acknowledgements, references and appendices are removed from the final list of texts in this corpus.

3. Key results

Key findings emerged as a result of the quantitative and qualitative analyses. First, expert mathematician writers exhibit control of and sensitivity to a wide range of academic clusters in comparison with graduate mathematician students who rely on a small number of highly recurrent strings. This gives further evidence to some previous findings which indicate that the successful use of lexical bundles hinges more on the academic maturity and expertise of the disciplinary writers, rather than on their native/non-native backgrounds (Haswell, 1991). Another finding is that a large number of lexical bundles in each writing group show a strong tendency to spread across registers, with the greatest proportion of the shared bundles found in the student mathematics discourse. Yet the other key finding relates to the structural and functional attributes of the bundles used by each writing groups. A great number of the lexical bundles in the student-produced texts are phrasal, while the majority of the bundles in the expert corpora are clausal, a finding that sheds doubts on the widely held assumption that as writers mature, their reliance on phrasal style become more clearly manifested (Pan et al., 2016).

Capitalizing on these results, there are several pedagogical implications for the teaching and learning of lexical bundles in the disciplines. Instructional intervention should focus not only on the forms that such lexical bundles display but also on the discourse functions that they serve. Given their perceptual saliency, the structural patterns of the academic bundles seem to be more conducive to learning and retention than the functional attributes which necessitate exposure to and familiarity with the contexts in which these bundles occur. In a form-focused situation, ESP/ EAP instructors and materials developers can use various techniques such as “highlighting” “underlining” or “bolding” to draw the attention of students and novice writers to the numerous structural types underlying bundles. The functions of the mathematical bundles can be instructionally fostered using labels discussed in this study or those alluded to by Cunningham (2017) and McGrath and Kuteeva (2012). These functional labels can be supplemented by examples derived from a specialized corpus, showing instances of occurrence as they are used by expert authors. Disciplinary textbooks and journal articles should also be incorporated into advanced ESP and EAP programs as writing models to emulate, for the reason that these texts are written by academically accomplished and profoundly experienced members whose publications undergo extensive peer-reviewing, editing and rewriting. A genre-based instruction targeting recurrent bundles could also help novice
mathematician writers accelerate transition from a student-styled use of language into more expert-styled one, a shift highlighted by Schleppegrell (2007) as posing a challenge to many aspiring novice mathematicians.

In the end, his study has shown that the mathematical discourse can be studied from a lexical bundle perspective, using a combination of quantitative and qualitative measures. It is hoped that the discussion above will enhance our understanding of how lexical bundles, defined as the most recurrent, widely dispersed n-grams in a discourse, are structurally and functionally used by mathematician writers at different levels of expertise.

4. References


Cortes, V. (2013). The purpose of this study is to: Connecting lexical bundles and moves in research article introductions. *Journal of English for Academic Purposes, 12*(1), 33–43.


Utilising corpus linguistic technologies: Quantifications in conversation analysis
Kazuki Hata (Tokyo City University, Japan)

This study proposes an interdisciplinary approach utilising conversation analysis (CA) in accordance with corpus linguistic (CL) techniques. CA stands as a micro-analytic approach designed to investigate the structure of social interaction (Sacks, Schegloff & Jefferson, 1974). Although CA is well-known as a qualitative approach, it inevitably incorporates a quantitative feature of research in that investigations of interactional organisation typically derive from the collections of samples (Schegloff, 1993). Thus, CA studies can certainly benefit from using CL approach as a methodological tool, as findings in corpora help identify the remarkable pattern(s) that can allow further investigations through a qualitative approach (Walsh et al., 2011). In the CA field, however, an application of CL methodology has not been well-explored in terms of how CL methods can be utilised in alignment with CA disciplines, and in particular, the significance of integration between computer-aided techniques and systematic micro-analytic procedures. The present study highlights that an application of CL can extend a CA in a way that manual analytical practices cannot achieve.

Although it has been continuously suggested that actual pragmatic features of language are not well-pursued by purely quantitative CL due to its inherent limitation: a lack of contextual information (Adolphs, 2008, p. 6–7; Mautner, 2007, p. 65–66; Sinclair, 2008; Widdowson, 2000), language corpora help studies across different disciplines in many ways by attaining significant numbers of language samples (Carter & McCarthy, 2006; Mautner, 2007, p. 54; McCarthy, 1998; Sinclair, 2008, p. 30). Spoken corpora offer tremendous insights from the authentic data, which becomes a reliable departure point for further case-by-case investigations. A utilisation of CL techniques thus has potential to help identify language samples to be analysed in the strict ‘data retrieval model’ (Leech, 1991, p. 20) for qualitative studies. That is, CL in this sense is a powerful tool to highlight language-in-use that is in accordance with interactional properties of target token in a particular context; investigating not only which word is used but also how the token is used and recognised by participants in a given interaction (Mautner, 2007, p. 54; Teubert, 2005, p. 8). Regarding a methodological combination, in particular between CL and CA, one prominent work is Walsh et al. (2011). They suggest a data-driven CL–CA approach, in which descriptive results (i.e. frequency lists) operate to be a good starting point for CA practices. They claim that the proposed CL–CA approach has a potential to provide detailed insights of interactions and bridge methodological limitation for each approach. However, one limitation in their current CL–CA approach is that the focus was considerably made on linguistic-centred aspects of interaction, providing partial proofs for the sequence organisational method of the participants and its systematic features on a turn-by-turn basis.

This work, utilising two corpora: the BNC Spoken Audio Sampler (Coleman et al., 2012) and Newcastle University Corpus of Academic Spoken English (NUCASE), tackles an application of CL techniques into strict ‘ethnomethodological’ CA perspectives on talk-in-interaction by suggesting two integrated approaches between CL and CA. The first approach considers the verbal production of guys in goal-
oriented interactions through the top-down CL–CA approach, wherein the research focus is characterised through a statistical result. A keyword list, generated from a comparison between the BNC and NUCASE data, identifies that a noun guys is predominantly found in NUCASE when compared to BNC Sampler, which becomes a base for a qualitative CA analysis. The following turn-by-turn exploration suggests that the token guys is used as an interactional resource to achieve a re-entry to the goal-oriented sequence from a diverted part of talk. In this top-down approach, the quantitative result can operate to identify a potentially distinctive and contextual feature of linguistic token used in social interaction; which can also be substitution for a manual ‘unmotivated-looking’ procedure in CA.

The second approach also deals with quantification in CA utilising CL techniques, yet is carried out in the opposite ‘bottom-up’ direction. This investigation is focused on the turn-final placement of erm (or um): a token placed at a possible turn transition point with ambiguous syntactic information (Jefferson, 1983; Local & Kelly, 1986). During the CA analytical process, an annotation is made on the target phenomenon by the researcher, which allows the comparison for particular social actions in line with the production of the target token between different datasets: ordinary and institution talk. Interestingly, the quantitative features of erm are differently seen between those contexts, implying that a turn-final erm is context-dependent and a simple classification (e.g. filler, conjunctonal or pragmatic marker) is insufficient to describe how the token is produced and recognised in social interaction. Unlike the top-down CL–CA procedure, this bottom-up approach can be reciprocal and can proceed to a further CA analysis when needed.

This study discusses the significance of quantification in CA studies made through applying CL techniques. Considering that organisation of talk in interaction is truly contextual, one would suggest that the linguistic patterns would not be able to be captured through truly quantitative procedures. Nevertheless, the utilisations of CL can provide clues for forthcoming case-by-case analyses, indicating the significant patterns for further investigations. Whereas the current project has not been complete, the results generated to date have already highlighted implications for such interdisciplinary approach.

References


The role of contexts of use in the study of learner language

Mick O’Donnell (Universidad Autónoma de Madrid, Spain)

This paper will explore the application of the notion of ‘context of use’ of linguistic forms in relation to learning a second language. By ‘context of use’ I mean the total environment in which a particular linguistic form (word, syntactic structure, or discourse structure) is appropriate. This includes not only Halliday’s ”Context of situation”, which is usually taken as global to a text as a whole, but also includes the referential context: the situation that the utterance is expressing.

Over the last few years, I have been studying the process of acquisition of syntactic features in learners of English using a syntactically parsed corpus of learner English from Spanish university students, graded by proficiency levels (e.g., O’Donnell 2013, 2015). One point that I have learnt from these studies is that one needs to distinguish two steps of the second language acquisition process: learning how to produce a form (syntactic competence) and learning when to produce the form (functional competence). In most cases, learners quickly learn how to produce a form, but then spend years mastering exactly in what contexts the given form should appropriately be used.

To give an example, learners of English from any L1 background have little trouble learning how to produce a syntactically correct noun phrase using a definite article such as the cats. However, they then spend years working out in which referential contexts the article is required or not. Similarly, most European languages have a constructed present perfect structure (I have eaten), and learners from these mother tongues can produce the English form without much trouble. However, the contexts of use of the present perfect differ subtly from language to language, and thus the learner is faced with the task of learning the contexts where the form is appropriate, expanding into new contexts of use not used by the mother tongue, or contracting away from some of the contexts used by the mother tongue.

In our prior study of errors made by Spanish learners of English (MacDonald et al, 2011), where we classified over 16,000 errors, we did not distinguish between formal and functional errors. For instance, determiner-noun-agreement errors (a formal error) were grouped under the same super-category with article-present-not-required errors (most of which are context-of-use errors). Also, lexical form errors (spelling errors, malformation errors) and lexical use errors (word-choice errors) were both together under vocabulary errors.

As a result of the insight that form and context of use errors present very different problems for language learners, we are reconfiguring our language studies with this form-function division as central to our error coding, and are placing more emphasis on identifying exactly which contexts of use are critical for each L1-L2 pair. In a sense, we are returning to the basic premises of Contrastive Analysis (Lado 1957), identifying the contexts of use of each syntactic form in the mother tongue and in English, and identifying the mismatches, which
are then potentially critical problems for the learners. For instance, in relation to the use of the present-perfect by Spanish learners of English, some of the relevant contexts of use are shown below. (Note: usage of this form in Latin America varies from that in Spain). One would expect a higher number of errors in the two contexts of use which result in different realisations for the two languages (marked by an “X”).

<table>
<thead>
<tr>
<th>Context</th>
<th>Spanish</th>
<th>English</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Same day past with no implied effect on present</td>
<td>At 5pm: He desayunado esta mañana</td>
<td>At 5pm: I ate breakfast this morning</td>
<td>X</td>
</tr>
<tr>
<td>Past happening with effect on present</td>
<td>Me he roto el brazo y no puedo trabajar</td>
<td>I’ve broken my arm and I cannot work</td>
<td></td>
</tr>
<tr>
<td>Achievements in still open period</td>
<td>Este año ya hemos construido 20 casas</td>
<td>We have built 20 houses so far this year</td>
<td></td>
</tr>
<tr>
<td>Life achievements</td>
<td>He vivido en cinco países distintos</td>
<td>I have lived in five different countries</td>
<td></td>
</tr>
<tr>
<td>Life achievements</td>
<td>(Es la primera vez que como sushi)</td>
<td>This is the first time that I’ve eaten sushi</td>
<td>X</td>
</tr>
</tbody>
</table>

For related work, see Ntouvis (2016) who explores the degree of misuse of tense-aspect combinations by EFL students, but this work does not identify in which contexts of use the errors are occurring.

One critique of Contrastive Analysis has been that not all inter-lingual differences lead to problems for the learner. However, our study is driven by our error analysis: we have identified those linguistic features which produce the most errors, and we then, within each of these areas, explore the reasons behind each error, identifying the contexts of use that are problematic for the learner. Firstly, we have identified the 10 most critical (frequent) error types in our corpus, and for each of these types, we are going through the errors, tagging them with more delicate features identifying the contexts of use in which the error is made. In this way, we are identifying the critical contexts of use in which most of these errors occur. Pedagogically speaking, our goal then becomes to make our learners aware of these problematic contexts of use. For instance, after re-coding our *article-present-not-required* errors, it became clear that the most frequent context for these errors related to generic reference (in particular, reference to generic abstract entities, and generic plural reference), and we have since put teaching material in place to make students aware of these two problematic contexts.

This paper will firstly present ideas behind using context of use in the exploration of the development of learner English. The talk will then outline the
results we have uncovered in relation to four studies either completed or in progress:

- **Article inclusion**: Dotti and O’Donnell (2012) examined those errors where articles were wrongly included in a noun phrase, coding more delicately in terms of the referential context, showing that the vast majority of errors involve generic plurals, generic abstract nouns, and percentages.

- **Quantification**: O’Donnell (2015) explored the uses of nominal quantification in the learner corpus, exploring which referential contexts produce most problems for the learners (e.g., ‘some’ should not be used in negated clauses).

- **Prepositions**: Gonzalez Díaz (in prep.) is examining cases of wrong preposition selection by Spanish learners of English, coding them in terms of the source preposition (back-translating the English into Spanish), to identify which Spanish prepositions prove the most problematic in translation into English. Later work will look more concretely at which more specific referential contexts produce these problems (e.g., Spanish “en” can be translated as “in”, “on”, “by”, “at” and “into” amongst others, and she believes that provision of the explicit contexts where each alternative is appropriate will help the learners).

- **Verb phrase errors**: Flores (in prep.) is studying to what degree errors in the verb phrase are errors of form (e.g., *I have shoot*) compared to context of use errors (e.g., use of present perfect where not appropriate).

**References**


Mixed Voices: A Corpus-based Study of Hungarian Parliamentary Debates about the European Union
Elena Valvason (University of Pavia, Italy)

1. Introduction

Emerging from a forty-year-long dictatorship, at the beginning of 1990s Hungary attempted a way back into the Western world: it revived its independence, it established democratic institutions, it opened its economic system to private property and to the global market. Hungary struggled to reconstruct its national identity in terms of peculiar Hungarianness and broader Europeanness, as a means to distinguish itself inside of a new comprehensive institution: the European Union (Kontler 2002: 20). Negotiations for EU-membership started on 31st March 1998 and ended on 12th April 2003, when the Hungarians voted to decide whether their nation should join the European Union. 83.8% of voters backed their country’s will to membership: this straightforward result followed a massive parliamentary discussion about the topic and it guaranteed Hungary the access to the EU (Fornaro 2006: 249).

Nevertheless, the low participation of the electorate, assessed around 45%, made two points clear: first, it showed that the Euro-sceptic discourse had already penetrated Hungarian politics; secondly, it suggested that the dichotomic character of Hungarian “national self”, alternating exclusive Hungarianness and full Europeanness, was not resolved (Romsics 2010: 566-567). But what were the Hungarian politicians’ perceptions of the European Union between 1998 and 2003? And what kind of identity did they build for their country in discourses about the EU? Tackling these issues, this paper presents the outcome of a corpus-based study of perceptions and identities in Hungarian parliamentary debates about the European Union.

2. Perceptions and identities: the study of semantic preference and semantic prosody

From a textual point of view perceptions and identities may be reflected in lexical choices clustering around the concepts of semantic preference and semantic prosody (Mautner 2007: 56), collocational phenomena respectively consisting in the relation between “a lemma or word-form and a set of semantically related words” (Stubbs 2001: 65) and the tendency of the same lemmas or word-forms to co-occur with items having a positive, negative or neutral prosody (Stubbs 1996: 176). Studies on Euro-sceptic discourse in Britain (Mautner 2000, Teubert 2000) and on European political discourse (with a particular focus on Hungarian in Krizsán 2011) showed that semantic preference and semantic prosody play a paramount role in sketching perceptions and identities, outlining the net of one’s ideological standpoint (Koller and Mautner 2004: 223). The former, in fact, spotlights the themes addressed by a speaker in relation to a pinpointed issue while the latter plots the same themes according to their positive, negative or neutral evaluative content (Partington 2004: 150-151).

3. Data and methodology: collocation analysis of the HUNPOL corpus

Aiming at the exploration of the Hungarian politicians’ attitudes towards the EU, the present research is based on a self-collected corpus of Hungarian parliamentary debates (the
HUNPOL corpus), that consists of 444 texts totalling to 344748 running words. The speeches were delivered by Hungarian MPs and external advisors at parliamentary sessions between 31st March 1998 and 12th April 2003. They have been collected querying the website of the Hungarian Parliament with the search term európai unió* (‘European Union*’) to guarantee that all the texts would be about the EU. A wildcard was employed at the end of the string owing to the agglutinative nature of the Hungarian language: Hungarian morphology is highly inflectional, resulting in the addition of derivational, grammatical and functional morphemes to the end of root items (Driussi 2012: 24). As a consequence, looking for európai unió only would have meant missing a list of forms like európai uniós ‘of the European Union’ or európai unióban ‘in the European Union’.

Semantic preference and semantic prosody were spotted in the corpus by means of collocation analysis. This analysis was carried out with GraphColl, a software that enables to visualise a word’s collocational pattern in the form of graphs (Brezina et al. 2015: 139). Collocates were extracted using Mutual Information (MI), an association measure that favours links between lexical words and that brings to the fore “highly specialised terms” (Evert 2008: 1238); MI minimum score was set to 5.0, after an empirical testing of the classical scores of 3.0 (Hunston 2002: 71) and 6.0 (Durrant and Doherty 2010: 145), and the minimum collocation frequency was fixed on 5 co-occurrences of node (i.e. the core item) and collocate. Although the analysis of only five instances of any word-form could not lead to generalisations, such a low frequency threshold was selected to avoid discarding relevant lemmas due to the inflectional nature of the Hungarian language. I decided to look for collocates inside a span of 3, 5 and 7 words to the left and to the right of the node in order to uncover a broader variety of repeated co-occurrences: from the ones closely modifying the search word to those shaping discourse. Combining quantitative and qualitative techniques, I explored the concordance lines of the first- and second-order collocates of my query term, so as to picture the whole collocational network of the concept I was interested in unveiling and to inform it with fine-grained details.

4. Analysis: swinging between collocates and concordances

To outline the Hungarian politicians’ perceptions of the ‘European Union’, I queried the HUNPOL corpus with the európai unió* string. Searching within a span of 3 words to the left and 3 to the right of the node, GraphColl retrieved 33 collocates, which I manually grouped according to their semantics in the “joining the European Union”, “requirements for EU-membership”, “membership”, “organisations other that the EU”, “being” and “Hungarianness” categories. As can be seen in table 1, the EU was essentially spoken about in relation to the possibility to achieve membership and to the tasks to fulfil to make it possible. No clear evaluative content emerged from mere collocation analysis but an in-depth perusal of the concordance lines showed that the European Union was viewed by some as an ‘opportunity’ (lehetőség) that was wished for and as a trigger for development (kívántunk alkalmazkodni ‘we desired to adapt’) whereas it was regarded by others as something to oppose (az európai uniós csatlakozás ellen ‘against joining the European Union’).

<table>
<thead>
<tr>
<th>Semantic category</th>
<th>Examples of collocates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joining the European Union</td>
<td>csatlakozás ‘joining’, csatlakozással ‘with joining’, küszöbén ‘approaching’</td>
</tr>
</tbody>
</table>
Requirements for membership

<table>
<thead>
<tr>
<th>Requirement</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>előírja '(it) requires', jogharmonizáció 'harmonisation of the law', tárgyalások 'negotiations'</td>
<td></td>
</tr>
<tr>
<td>Membership</td>
<td>országáiban 'in its countries', tagállamok 'member-states', tagságunk 'our membership'</td>
</tr>
<tr>
<td>Organisations other than the EU</td>
<td>nato 'NATO'</td>
</tr>
<tr>
<td>Being</td>
<td>leszünk '(we) will be', történő 'happening', való 'existing'</td>
</tr>
<tr>
<td>Hungarianness</td>
<td>hazánk 'our country', magyarország 'Hungary', magyarországnak 'to Hungary'</td>
</tr>
</tbody>
</table>

Table 1. Exemplary collocates of *európai unió* – MI(5.0), 3L-3R, C5-NC5

Expanding the collocation span to 5 and to 7 words to the left and to the right of *európai unió*, the collocates increased to 41 and 81 respectively. In the former case, they simply added new items to the aforementioned six categories (e.g. megállapodás ‘agreement’ and támogatások ‘funding’) while in the latter they provided new perspectives on the EU, encoded in lexical (for instance, teszünk ‘(we) do’) and in stylistic features (as in márpedig ‘well’ or szempontjából ‘from the point of view of…’).

The Hungarian identity started being characterised in the “Hungarianness” grouping through word-forms like ‘to Hungary’ (*magyarországnak*) and ‘our country’ (*hazánk*) and it was further explored via the collocational networks of *magyarország* (*Hungary*) and *magyarok* (*Hungarians*). With the 3, 5 and 7 collocation spans, ‘Hungary’ counted 13, 28 and 68 collocates and it appeared in terms of oppositions: it was active in wishing to join the EU (e.g. *kíván* ‘desire’, csatlakozási ‘of joining’) but still signalling its own borders (for example in *határait* ‘its borders’, *tartózkodás* ‘stay’, *külföldiek* ‘foreigners’); it was or it could have been successful in Europe (*Magyarország lehet sikeres* ‘Hungary can be successful’) but it was still underdeveloped to some extent ([a kormány] *ne feledkezzen el a korábban méltatlanul elhanyagolt vidéki Magyarországról sem* ‘the government should not forget the Hungarian countryside, that has been unfairly neglected’).

Through the 9 collocates of ‘Hungarians’, also the Hungarian people were portrayed in dichotomic terms: they differentiated ‘Hungarians living in their homeland’ (*anyaországon belül élő magyarok*) from those staying ‘outside the national borders’ (*határon túli*) or more precisely ‘dwelling in the neighbouring countries’ (*szomszédos államokban élő*) and needing to be protected by laws (*A szomszédos államokban élő magyarokról szóló törvény megalkotásának a vége felé közeledünk* ‘We are approaching the completion of the law regarding the Hungarians living in the neighbouring countries’)

5. Conclusion

The study showed that, in the years preceding Hungary’s joining the European Union, the Hungarian politicians referred to the EU with mixed voices, alternating positive stances, claiming the need to join the EU to develop further and faster, and negative standpoints, warning that the EU-membership could cause a loss of national strength. The same doubts appeared even in the characterisation of the Hungarian identity. Hungarians pictured themselves as people in between: on the one hand, they hoped to enter the European Union and achieve full Europeanness, but on the other they convincedly recalled their Hungarianness in the historical division of their people across the country’s borders.
References


A Parallel Corpus-Based Study of Chinese Arabic Verb Phrase Alignment
Alaa Mamdouh Akef, Yingying Wang and Erhong Yang (Beijing Language and Culture University, China)

1. Introduction

The Bilingual Alignment Corpus is composed of five categories that vary according to their alignment units: paragraph, sentence, lexical, chunk, and phrase alignments. Over the years, researchers have been committed to the phrase alignment of bilingual corpora based solely on the syntax and lexical alignments.

In this paper, we used an "Analysis - Analysis - Matching" alignment strategy to extract the Arabic-Chinese alignment pairs. First, we used a lexical alignment for a parallel Arabic-Chinese corpus, and in accordance with this result, we extracted the phrase alignment pairs based on syntactic analysis and we built a parallel Arabic-Chinese phrase corpus. Then, we selected Arabic verb phrases as the object of our study, afterwards we examined the Arabic verb phrases and their corresponding Chinese phrases’ internal structures and after discovery of the Arabic-Chinese verb phrase alignment rules, we summarized them; which built translation rules for instance-based and rule-based machine translation systems.

2. Experiment

First, 7,125 Chinese-Arabic aligned sentences with 827,500 words were drawn from UN official transcripts; the Arabic data included 7,125 sentences and 827,500 words, and the Chinese data included 7,523 sentences and 265,427 characters. The total of sentence pairs after sentence alignment was 7,125. Second, the errors of the parallel corpus were simply modified manually, including typos, Arabic collocation, punctuation and titles.

The experiment was carried out as follows:
This paper used the word alignment tool GIZA ++ (Och & Ney, 2003) to complete the two-direction (Arabic-Chinese and Chinese-Arabic) lexical alignment. The GIZA ++ tool process specifies that each source language word corresponds to only a single target language word. When the source language word does not correspond to the target language word, it is assumed to be aligned to a null word. We observed the existence of null words, which are caused by the huge lexical, word order, and syntactical differences between the Arabic and Chinese linguistic systems.

The Stanford parser arabicFactored was used for the Arabic data syntax parsing (Green & Manning, 2010), while xinhuaFactoredSegmenting was used for the syntactic parsing of the Chinese data (Levy & Manning, 2003). During the analysis of the parsing results, it was found that both Arabic and Chinese had parsing errors, such as the tagging of a blank element with some part of speech or syntactic element, e.g. $(NP \ (NN ))$. After removal of the blank elements within the sentence pairs, 6,387 sentence pairs were left.

Based on the syntactic analysis and word alignment results, this paper used the "Analysis - Analysis - Matching" strategy to write a program to automatically extract Arabic-Chinese aligned phrase pairs. In 1992, Kaji first used this method for Japanese-English bilingual phrase alignment, and afterwards it was applied to other language pairs. In this paper, a total of 19,586 alignment phrase pairs were extracted from 6,387 Chinese sentences, and were used as data to build the Arabic-Chinese parallel phrase bank.

Arabic is a VSO language, and this experiment’s results also show that VSO word order accounts for 82.7% of the whole Arabic corpus, leading us to choose Arabic verbal phrase as our research object. After deleting the 1,088 phrases which had syntactic parsed errors, we had 4,833 verbal phrase pairs,
on which the basis of the internal composition of the Arabic verbal phrase and its corresponding Chinese phrase structures were summarized. Arabic verbal phrase and Chinese phrase alignment rules were also summarized.

3. Results and analysis

The syntax parsing errors for the above-mentioned 1,088 Arabic verb phrases mainly included tagging prepositions, nouns, Arabic ann particles, function words, demonstratives, pronouns, and other words as verbs. The author describes this question in more detail in another paper about Arabic-Chinese phrases (Akef, Yang, & Wang, 2016).

3.1 The internal composition of Arabic verb phrases

The three main types of verb phrases are verb + noun phrase (V + NP), verb + prepositional phrase (V + PP), and function word + verb (RP + V). The tense of the verbs in each type can vary.

Table 1 The Internal Composition of the Arabic Verb Phrase

<table>
<thead>
<tr>
<th>Internal composition</th>
<th>Examples</th>
<th>Number</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>VP-&gt;VBD+NP</td>
<td>(VP (VBD (التجارة (متارت (NP (DTNN (الخارجية)))</td>
<td>3,032</td>
<td>69.1%</td>
</tr>
<tr>
<td>VP-&gt;VBP+NP</td>
<td>VP (VBP (السلام))</td>
<td>869</td>
<td>19.8%</td>
</tr>
<tr>
<td>VP-&gt;RP+VBP</td>
<td>VP (PRT (RP (لا (VBP (بستغني (NP (NN (واستعترر (NP (DTNN (العالم))))))))</td>
<td>136</td>
<td>3.1%</td>
</tr>
<tr>
<td>VP-&gt;VBP+PP</td>
<td>VP (VBP (نحافظ (PP (IN (NP (CD 100) (PUNC %))))))</td>
<td>61</td>
<td>1.4%</td>
</tr>
<tr>
<td>VP-&gt;VBD+PP</td>
<td>VP (VBD (من (NP (IN (أداة (NP (NN (دارة (NP (DTNNS (المؤسسات))))))))</td>
<td>61</td>
<td>1.4%</td>
</tr>
<tr>
<td>VP-&gt;VBN+PP</td>
<td>VP (VBN (شتركت (PP (IN (NP (DTNN (لإنتاجة (التمار)))))</td>
<td>61</td>
<td>1.4%</td>
</tr>
<tr>
<td>VP-&gt;VBN+NP</td>
<td>VP (VBN (حصة (NP (NN (وابصارر (NP (NN (نانوتهم (JJ (وارداتها))))))))</td>
<td>61</td>
<td>1.4%</td>
</tr>
<tr>
<td>VP-&gt;VN+PP</td>
<td>VP (VN (بال (PP (IN (النجاج (NP (NN (وابصاراتهم (JJ (وحدده (شاقه (JJ (وارداتها))))))))))</td>
<td>75</td>
<td>1.7%</td>
</tr>
<tr>
<td>VP-&gt;VBD+ADJP</td>
<td>(ROOT (S (VP (VBD (وقوع (ADJP (JJR (PP (IN (من (NP (CD 150) (NP (NN (بلدا))))))))</td>
<td>32</td>
<td>0.7%</td>
</tr>
</tbody>
</table>

Total                  |                                                                        | 4,388  |         |
3.2 Chinese Phrase Structure with Verb Phrase Alignment

Through the analysis of the structure of the corresponding Chinese phrases for the three types of internal composition of the Arabic verb phrases, we may find that the structure of the Chinese phrase is different for each one, as follows:

1) Verb + noun phrase (V + NP)

When the internal structure of the verb phrase is verb + noun phrase, there are five corresponding Chinese phrase structures, as shown in the following table:

<table>
<thead>
<tr>
<th>Structure of Chinese Phrase</th>
<th>Arabic verbal phrase</th>
<th>Chinese phrase</th>
</tr>
</thead>
<tbody>
<tr>
<td>VV+NP+VP</td>
<td>(VP (VBP (بسط) (PP (IN (على) (NP (DTNN (المرض) (DTJJ值得)) ومكن)))</td>
<td>(VP (VP (VV 使) (NP (NN 疾病)) (VP (VV 得到) (NP (NN 控制)))))</td>
</tr>
<tr>
<td>VV+VP</td>
<td>(VP (VBP (تساعد) (PP (IN (على) (NP (NN (توسع) (NP (DTNN (الهوائية) (DTJJ الرئيسية)))</td>
<td>(VP (VV 帮助) (VP (VV 扩张) (NP (NP (NN 肺部))))</td>
</tr>
<tr>
<td>VV+NN</td>
<td>(VP (VBP (يؤدي) (PP (IN (إلى) (NP (NP (NN (تضيق) (NP (DTNN (التنفسية) (DTJJ المسالك)))</td>
<td>(VP (VV 导致) (NP (NN 气道狭窄))</td>
</tr>
<tr>
<td>V+PP</td>
<td>(VP (VBP (يظهر) (PP (IN (في) (NP (NOUN_QUANT (جميع) (NP (DTNN (البلدان)))</td>
<td>(NP (PN 它)) (VP (VV 发生) (PP (P 在) (NP (NP (NN 国家)))))</td>
</tr>
<tr>
<td>NP+ADJP</td>
<td>(VP (VBP (يختلف) (PP (IN (في) (NP (NP (NN (شدة) (JJ 其重))) وتوترها))</td>
<td>(NP (PN 其) (ADJP (JJ 严 重) (NP (NN 程度)))))</td>
</tr>
</tbody>
</table>

2) Verb + prepositional phrase (V+PP)

When the internal structure of the verb phrase is verb + prepositional phrase, there are five corresponding Chinese phrase structures, as shown in the following table:
### Table 3 Arabic V+PP corresponding Chinese phrase structure

<table>
<thead>
<tr>
<th>Structure of Chinese Phrase</th>
<th>Arabic Verbal Phrase</th>
<th>Chinese Phrase</th>
</tr>
</thead>
<tbody>
<tr>
<td>NP+VV+NP</td>
<td>(VP (VBP يضم NP (NNS منظمات JJ الوطنية JJ ) ) )</td>
<td>(NP (PN 它)) (VP (VC 是) (NP (DNP NP (NP (QP (CD 一点点)) (CLP (M 个))) (NP (NN 国家) (CC 和) (NN 国际)) (NN 组织)))</td>
</tr>
<tr>
<td>PP+NP</td>
<td>(VP (VBP يصيب NP (DTNN الرجال ) (DTJJ ) ) )</td>
<td>(VP (PP 对) (NP (DNP NP (NN 男性) (CC 和) (NN 女性)) (DEG 的)) (NP (NN 影响)))</td>
</tr>
<tr>
<td>VV+NP</td>
<td>(VP (VBP يعوق NP (NN عملية NP (DTNN التنفس ) ) ) )</td>
<td>(VP (VV 减缓) (NP (DNP NP (NN 疾病) (DEG 的)) (NP (NN 发展)))) (PU 。)</td>
</tr>
<tr>
<td>NP+VP</td>
<td>(VP (VBP يتخلى NP (NNS منظمة NP (DTNN الصحة ) ) ) )</td>
<td>(NP (NN 世界) (NN 组织)) (VP (ADVP AD 还)) (VP (VV 领导))</td>
</tr>
<tr>
<td>VP</td>
<td>(VP (VBP يركز NP (PUNC ، ) (NP (NN تحديثا ) ) ) )</td>
<td>(VP (ADVP AD 具体)) (VP (MSP 来) (VP (VV 说))))</td>
</tr>
</tbody>
</table>

### 3) Function word + verb (RP+V)

Function words are a very unique language phenomenon within the Arabic language, not having a specific meaning by themselves, but playing a very important part of grammatical function.

When the internal structure of the verb phrase is function word + verb, there are five corresponding Chinese phrase structures, as shown in the following table:

<table>
<thead>
<tr>
<th>Structure of Chinese Phrase</th>
<th>Arabic verbal phrase</th>
<th>Chinese phrase</th>
</tr>
</thead>
<tbody>
<tr>
<td>VP</td>
<td>(VP (RP لا (VBP ) (VBP بريقين ) (VP (PRT (RP ) (VBP قد ) (VP (NN ظهر ) (NP (DTNN الفوارة ) ) ) ) ) )</td>
<td>(VP (VA 不足))</td>
</tr>
<tr>
<td>NP</td>
<td>(VP (ADJP ل (VBP ) (VP (ADVP AD 直接)) (NP (NN 并发症)) ) )</td>
<td>(NP (NN 疾病))</td>
</tr>
<tr>
<td>IP</td>
<td>(VP (ADVP AD 后)) (VP (VV 利用)))</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(NP (NN 世界) (NN 组织)) (VP (ADVP AD 还)) (VP (VV 领导))</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(IP (NP (NN 废水)) (VP (ADVP AD 再)) (VP (VV 说)))</td>
<td></td>
</tr>
</tbody>
</table>
4. Conclusion

The internal composition of verb phrases in Arabic and the structure of their corresponding Chinese phrases were measured with a parallel Arabic-Chinese phrase bank, taking 4,388 Arabic verb phrases as objects of the study composition with regards to the verb phrases. We found that apart from verb phrases, when translating Arabic verb phrases into Chinese, noun phrases and prepositional phrases can be used as well, and every type of Arabic verb phrase can correspond to various Chinese phrase structures. The summarized correspondences obtained from an authentic corpus can be used as phrase-alignment rules in an Arabic-Chinese machine translation, and can be applied to instance-based or rule-based machine translation systems.

References


Sentence-internal capitalization of nouns is a distinctive characteristic of the German spelling system. Its diachronic emergence in the 16th and 17th centuries has received much attention in recent decades (e.g. Kämpfert 1980, Moulin 1990). Most importantly, a fairly comprehensive corpus study on the basis of printed texts has contributed significantly to our understanding of this phenomenon (Bergmann & Ne- rius 1998). However, some important questions regarding the emergence of sentence-internal capitalization remain unanswered. Firstly, previous studies have only provided monofactorial analyses, focusing on a small number of semantic and pragmatic factors such as animacy and reverence. The interaction of these different factors, as well as additional ones like word frequency, has not been taken into account so far. Secondly, previous research has largely focused exclusively on printed texts (with the exception of Moulin 1990, who only investigates the writings of one author). However, it seems promising to take handwritten texts into account as well, as they are produced in a more spontaneous way as compared to printed texts and can therefore provide a glimpse into (quasi-) spontaneous language production. Handwritten texts might therefore be particularly well-suited for investigating the impact of cognitive, semantic, and syntactic factors in the use of sentence-internal capitalization.

In order to investigate the emergence of capitalization in Early New High German handwritten texts in more detail, we compiled a corpus of witch interrogation protocols, based on the edition by Macha et al. (2005), which comprises text samples from six dialect areas and covers the time span from 1570 to 1653. This period marks the incipient stages of sentence-internal capitalization. Consequently, there is a large amount of variation in the texts, which allows for investigating the factors that boost the use of sentence-internal capital letters. Fig. 1 shows the regional distribution of the texts. The six different shapes used for the individual datapoints represent the language areas assumed by Macha et al. Using digital facsimiles, we double-checked for samples of each text that the edition faithfully represents the handwritten originals. In sum, the corpus contains about 90,000 tokens. A multi-layer annotation was implemented, sentence boundaries were tagged according to a complex set of criteria (as punctuation is sparse and unreliable in many of the texts; cf. Szczepaniak & Barteld 2016 for details), and all tokens tagged as nouns were annotated for animacy using a fine-grained animacy hierarchy, as established hierarchies like the one proposed by Zaenen et al. (2004) proved inadequate for the given text type (cf. Barteld, Hartmann & Szczepaniak 2016).
Drawing on these data, we fit a binomial mixed-effects regression model using R (R Core Team 2016) and lme4 (Bates et al. 2015). Two variables were entered as fixed effects into the model: Animacy and Frequency. The fine-grained animacy annotation was broken down to a five-way distinction (“abstract”, “concrete”, “animal”, “human”, “superhuman”, the latter including terms like God, devil, demon, which are fairly frequent in the given text type), as a too fine-grained variable would lead to a rank-deficient model. The idea behind adding frequency as a factor is that we can expect more frequent words to have a fairly fixed graphemic gestalt, while we would expect more variation for low-frequency items (cf. e.g. Kapatsinski 2010). The token frequency of each type was obtained from the corpus itself. An alternative would have been to use another, larger corpus, but we decided against this option for two reasons: Firstly, there is currently no significantly larger corpus covering the time period in question, and secondly, the vocabulary used in the witch interrogation protocols is partly very specific to this text type. In addition, random intercepts for Lemma and Protocol as well as random slopes for the fixed effect of Animacy were added. Tab. 1 shows the results of the model.

|                     | Estimate | Std. Error | z value | Pr(>|z|) |
|---------------------|----------|------------|---------|----------|
| (Intercept)         | -2.01    | 0.25       | -8.05   | <0.001***|
| Animacy-concrete    | 1.27     | 0.27       | 4.62    | <0.001***|
| Animacy-animal      | 1.7      | 0.59       | 2.87    | <0.001***|
| Animacy-human       | 1.82     | 0.38       | 4.81    | <0.001***|
| Animacy-superhuman  | 3.79     | 1.28       | 2.97    | <0.001***|
| log_{10}(Freq)      | 0.39     | 0.2        | 1.9     | 0.06     |
| Animacy-conc×log_{10}(Freq) | -0.85 | 0.26 | -3.32 | <0.001*** |
| Animacy-anim×log_{10}(Freq) | -0.17 | 0.68 | -0.25 | 0.8      |
| Animacy-hum×log_{10}(Freq) | -0.35 | 0.32 | -1.11 | 0.27     |
| Animacy-sup×log_{10}(Freq) | -1.79 | 0.76 | -2.36 | 0.02     | *

Tab. 1: Coefficients of the fixed effects in the mixed model.

According to a log-likelihood test, the use of the random intercepts and slopes mentioned above is warranted as the model performs significantly better than models without the random effects for Lemma and Protocol. Importantly, the model also performs better than null models without Animacy ($\chi^2=327$, df=22, $p<0.001$***), or without the interaction term between Animacy and Frequency ($\chi^2=15.7$, df=4, $p=0.003$**). The index of concordance $C$, which assesses how well the model predicts the data (Baayen 2008: 204), indicates a very good fit ($C=0.94$, considered “outstanding discrimination” by Hosmer & Lemeshow 2000: 162). All Variance Inflation Factors (VIFs) – which are used to check for potential multicollinearity – are below the threshold of 5, which is often mentioned as a rule of thumb in the statistical literature (Levshina 2015: 272). In sum, the present study confirms the results which Barteld et al. (2016) have obtained on the basis of a much smaller and less representative sample: While the effect of animacy shown in earlier studies is very clearly substantiated, animacy alone cannot explain the variation in the data. We predicted that token frequency might have an effect on capitalization, but according to the results of the model, this effect is not straightforward. In general, highly frequent items seem more prone towards capitalization than low-frequency items. However, there is much lexeme-specific variation among high-frequency items which can partly
be explained in terms of animacy. For instance, *Mann* ‘man’ is capitalized in 69 out of 119 cases (81%) and *Gott* ‘god’ in 174 out of 189 (92%), while only 18 out of 117 occurrences of *Tag* ‘day’ (15%) are capitalized. However, some possibly important factors have not been captured in the model yet. Aspects that need further investigation include pragmatic motivations and syntactic factors in the early use of sentence-internal capitalization. For example, it is striking that in a sample of 18 texts, terms denoting women are significantly less often capitalized than terms denoting men (Fisher exact test: $p<0.001$, odds ratio=3.01), which points to an influence of cultural factors. In addition, future research could extend the focus from bare nouns to entire NPs, taking their internal structure into account.

Acknowledgments

The research reported on in this paper was funded by the German Research Foundation (2013–2015: SZ 280/2-1, KO 909/12-1; 2016–2018: SZ 280/2-3).

References


& D. Byron (Eds.), DiscAnnotation ’04 (pp. 118–125). Stroudsburg, PA: Association for Computational Linguistics.
Badgers, hedgehogs and squirrels: a diachronic corpus-assisted discourse study of British wildlife in the news
Emma McClaughlin (Lancaster University, UK)

Introduction and rationale

A recent Leverhulme-funded project (2013 – 2016) investigated the discursive representation of animals in contemporary Britain across a range of spoken and written discourse types. My own research attached to this project uses a modern diachronic corpus-assisted discourse analytical (MD-CADS) approach (Partington, 2010) to investigate how a small number of key wildlife species in Britain – badgers, hedgehogs and squirrels – are represented in a diachronic corpus of news texts published over a 220-year period.

The “wildlife” classification represents a particular type of social relationship between humans and the animals belonging to it. Wildlife species are not the “property” of humans, unlike domesticated species bred to fulfil a human requirement; for food, products or companionship, for example. It has been established that human perception of animals is influenced by social, political and cultural factors (Kean, 2001) and this is reflected in the language used to represent them, as demonstrated by Arran Stibbe’s work in particular (2001, 2003, 2006; and 2012). Human experience of animals in Britain has undergone major shifts since the industrial revolution; factors such as urbanisation, technological and agricultural advances – all consequences of industrial advancement – have changed the landscape of Britain, impacting on the way people interact with nature and wildlife (and Lovegrove, 2007; see Ritvo, 1987). The status of wildlife species, including squirrels, badgers and hedgehogs, has been subject to change historically as a result of these and other external factors which cause humans to perceive and manage their relationships with them differently. Importantly, the roles animals play in human society are imposed by humans – animals cannot self-identify as belonging to a particular social group, neither can they subscribe to or resist human discourses. If, as it has been argued, linguistic representations of human social actors and groups historically can have implications for how these people are perceived in present day society (see, for example, Baker, 2014) then it is worthwhile to study the role of language in shaping human relationships with animals in historical society, particularly since people tend to frame present issues with reference to the past and (re-)interpret the past with reference to current understanding (see Carr, 1990).

The intended outcomes for this study at the outset were to establish:

(1) the role of language in defining human-wildlife relationships in British cultural history

---

which linguistic factors – motivated by external political, social and cultural influences – have contributed to the (un)sustainability of human-animal relationships in the past. 

(3) the extent to which the findings can be used to inform understanding of future linguistic representations of wildlife

Data and methods

As well as identifying changes in linguistic variation, MD-CADS can be used to study the influence of social, political and cultural perspectives on language change (see Corpora special edition edited by Partington, 2010; and Partington, 2012). This paper discusses the impact of such influences on news discourse about squirrels, badgers and hedgehogs across the late-modern period of English. Unlike traditional MD-CADS, I did not use temporally parallel corpora for diachronic language comparison but rather a (relatively) continuous dataset gathered specifically for this study. Three corpora of news texts (letters and articles) gathered from the Times Digital Archive were constructed – one each for badgers (714 texts and 372172 tokens), hedgehogs (455 texts, 296516 tokens), and squirrels (including both the red and grey species) (696 texts, 328140 tokens). The corpora comprised all texts containing at least one mention of the animals under investigation published in The Times between 1785 and 2005.

The resulting diachronic distribution and the limited size of the corpora had a bearing on the methods available for carrying out a diachronic analysis. Each of the three corpora were segmented into distinct periods for diachronic comparison guided by the waves, peaks and troughs (WPT) statistical method (Gabrielatos, McEnery, Diggle, & Baker, 2012) using Brezina’s (2014-2015) “Change over time” tool (from the “Statistics in corpus linguistics toolbox” provided by CASS at Lancaster University). This analysis allows the researcher to make an informed decision about corpus segmentation points given peaks and troughs in the diachronic distribution of texts; these peaks and troughs reflect the frequency difference between articles published in each year, compared with the preceding year.

I utilised a multi-perspective analytic method comprising six individual corpus analyses for each animal (“red squirrel(s)”, “grey squirrel(s)”, “badger(s)” and “hedgehog(s)”) across the respective corpora; the analysis allowed me to gain a broad sense of the dominant representations of each of the animals under investigation, as well as the ways in which such representations have changed over time. A “keywords by corpus” analysis identified keywords for each of the corpora in turn, using the other two corpora as a single reference corpus. The remaining five analyses explored the diachronic aspect of the data by comparing texts contained within individual time periods with the remainder of the relevant corpus. I applied the diachronic categorisation proposed by McEnery and Baker (2015) in their diachronic collocates analysis – initiating, transient, consistent and terminating – to the corpus outputs from three analyses: diachronic collocates, diachronic keywords, and diachronic modifiers. Finally, I examined the behaviour of two clusters that were present across each of the corpora, over time: [species name]+BE; and of (the) [species name].

This multi-perspective analytical method highlighted a number of broad themes and sub-themes in the linguistic representation of squirrels, badgers and
hedgehogs; for example, I found language relating to the themes of *geography, places and spaces; animal qualities, states and attributes; and actions, pursuits and behaviours*. I then selected three major sub-themes identified in this process (see Findings below for details) for investigation using more detailed corpus-assisted discourse analysis. Each of these selected themes can be found across all three corpora and can be seen to change over time in response to text external factors. For each selection, I ensured that there was both enough data to examine and that this data was suited to diachronic analysis.

**Findings**

I will present key findings from this study, discussing them in the context of the broader social, cultural and political history of Britain in line with the discourse historical approach (DHA) (Reisigl & Wodak, 2009). These findings are linked to three major sub-themes: (i) *life cycle and health*, which includes language about animal age, the seasons, and disease; (ii) *origin and nationality*, which includes language relating to spatial distribution, the presence and absence, and the expansion and decline of species; and (iii) *human actions and pursuits*, which includes language about defence, protection and promotion of wildlife populations as well as the control, killing or otherwise harming of wildlife for country sports, population control (culling) and accidental death by machines (e.g. cars). I discuss patterns of change and continuity in the language about these animals in relation to these three sub-themes and the extent to which these patterns are consistent with changing human practices and attitudes. Specifically, I demonstrate in this paper the ways in which *anthropocentrism* (the tendency to interpret reality on the basis of human values) – an overarching topic common to all three corpora and all three selected sub-themes – is evident in the language used to represent squirrels, badgers and hedgehogs, and the behaviour of this topic over time in response to social, cultural and political external factors.

In short, I found that squirrels, badgers and hedgehogs are discussed in relation to human interests in news discourse. This might be because they are considered to cause financial or material losses; they are used for human amusement or hinder people's enjoyment of another species; or, more indirectly, because parallels are drawn between humans and animals, to the extent that a wildlife species may even stand as a proxy for human social groups. I conclude that by examining the ways in which wildlife species have been represented in the past, as well as the ways in which these representations respond to text-external social, cultural and political factors, we might better be able to understand human responses to wildlife-related issues, both current and beyond.

**References**


Revisiting the idiom principle through the lens of an agglutinating language: A corpus-based description of adjective-noun collocations in Turkish and English
Doğuş Can Öksüz and Vaclav Brezina (Lancaster University, UK)

Most research into formulaic language has been limited to a narrow set of languages particularly English (Durrant, 2013). For example, Sinclair’s (1991) “idiom principle”, which states that a language user has large number of available semi-preconstructed phrases that constitute single choices during processing, - has only rarely been applied to highly inflected, agglutinative languages such as Turkish. For this reason, our knowledge on the role they play in formal, functional, pragmatics and processing aspects, has been restricted to the narrow set of languages and the status of formulaicity as a property of language as such has not sufficiently established (Durrant, 2013). Biber (2009) suggested that agglutinating languages such as Turkish and Finnish are particularly interesting candidates for investigating the scope and status of formulaicity as a property of language. In this paper, therefore, we demonstrate that the formulaicity in agglutinating languages such as Turkish is different from the formulaicity in non-agglutinating languages. One of the reasons is that the rich morphology of an agglutinating language such as Turkish, affects the frequency of occurrence and syntagmatic associations between lexical items.

This study, following a frequency-based tradition of collocation research (Evert, 2008, Hoey, 2005; Sinclair 2004), is a corpus-based description of two-word adjective-noun collocations in Turkish and English. Corpora chosen for this study represent the input that language users experience on a daily base. The Turkish National Corpus (TNC) is a written, and general corpus of Turkish with a size of 47,641,688 tokens. It is a collection of 4438 different text samples, representing 9 domains and 34 different genres, written in between 1990-2009. The British National Corpus (BNC) XML edition was used as a comparison corpus. To extract the adjective-noun collocations, frequency bands were established making use of the TNC and the BNC word frequency lists. For establishing the bands, the frequency scale of the nouns in the TNC and the BNC were considered and five node words were selected from high and mid-frequency bands. Only the nouns were selected as node words to search for adjective collocates. Using the raw frequency scores, the most frequent four adjective collocates of each node words, a total of forty collocations were extracted for further analyses on two-word collocations’ frequency of occurrence and collocational strength. The adjective-noun collocations were chosen as focus of the investigation for two reasons: First, nouns within adjective-noun collocations can be inflected with various types of suffixes including case marking, plural and instrumental in Turkish, and thus it is possible to observe the influence of agglutination on the collocability of adjectives and nouns. Second, they occur in a certain syntactic order in which adjectives precede the nouns in both Turkish and English, as observed in the concordances lines; hence they should be
fully comparable for both strength and directions of the syntagmatic associations in Turkish and English.

For the frequency comparisons, raw frequency scores of adjective-noun collocations in the BNC and the TNC were relativized to per million words. This allows a comparison of how many times a two-word adjective-noun collocation in Turkish is likely to occur per million words against its equivalent in English. Besides the frequency comparison, this study focused on the collocational strength, the association between the node words and the collocates as measured by Mutual information (MI), Log Dice and Delta P metrics. The collocational relationship is quite complex and no single measure of association can capture the full complexity of this relationship (Brezina, McEnery, & Wattam, 2015). Therefore, this research used three different corpus derived measures of associations to investigate the collocational strength. MI scores were used to measure the rare exclusivity, that is, MI score tends to highlight the relatively infrequent words with low co-occurrence frequency (Evert, 2008, Manning & Schütze, 1999). Thus, another measure of association, which is neutral to the low frequency of occurrence, was needed. In this regard, Log Dice provides both an accurate measure of association and easily interpretable scale of scores (Rychly, 2008). MI and Log dice measures consider collocational strength as necessarily symmetrical. For this reason, Delta P was used as a measure of probability that takes directionality of the collocational strength into account. In this study, Delta P was used as a measure of association alongside MI and Log Dice measures to investigate whether adjectives are more predictive of the following nouns or the nouns are more predictive of the preceding adjectives (Gries, 2013).

Lemmatisation of the nominal inflections within adjective-noun collocations to abstract away from the complex morphology of Turkish was a natural step. For calculating the lemmatised forms of the collocations’ relative collocate frequency and association scores in Turkish and English, the raw frequencies of the node words’ all of the inflected forms and the collocations’ all of the inflected forms were identified. The frequency sums of the inflected forms of the node words were taken as frequency of the node in the whole corpus and the frequency sums of the all inflected forms of the collocations were taken as frequency of the collocation in the collocation window. Thus, the relative collocate frequency and association scores, as measured by MI, Log Dice and Delta P, were calculated for both collocations’ inflected and uninflected forms in Turkish English. After calculating the relative frequency and association scores for each inflected forms, the same measures were calculated for the lemmatised form of the collocations.

Overall, this study made possible to observe the relative collocate frequency and association scores of each inflected forms of the adjective-noun collocations. Furthermore, it was possible to make a comparison between the lemmatised and unlemmatised forms of the adjective-noun collocations for frequency of occurrence and associations in Turkish and in English. The data revealed that (55%) of high-frequency band, and (50%) of mid-frequency band collocations reach higher unlemmatised relative collocate frequency scores in English than their equivalents in Turkish. Lemmatised relative collocate frequency scores indicate that (70%) of high-frequency
band and (75%) of mid-frequency band collocations reach higher scores in Turkish than their equivalents in English. This may be viewed as a surprising finding considering the fact that lemmatised forms of the collocations in English and Turkish are functionally equivalent. Unlike relative collocate frequency scores, the adjective-noun collocations reach predominantly higher unlemmatised and lemmatised MI scores in English than in Turkish. Unlemmatised MI scores indicate that only (20%) of high-frequency band and (40%) of mid-frequency band adjective-noun collocations reach higher MI scores in Turkish than their equivalents in English. Similar to unlemmatised MI scores, only (25%) of high-frequency band, and (40%) of mid-frequency band adjective-noun collocations reach higher lemmatised MI scores in Turkish. To conclude, adjective-noun collocations in Turkish tend to be more frequent, but weaker associated than their equivalents in English. This might have implications that speakers of Turkish and English might need to process adjective-noun collocations differently in their respective L1s.

References


Investigating mentoring in a teacher development programme from a corpus-based hermeneutic-phenomenological perspective
Marilisa Shimazumi (Faculdade Cultura Inglesa São Paulo, Brazil)

1. Introduction

This paper is the result of a doctoral research project which aimed at describing and interpreting the phenomenon of mentoring from a hermeneutic-phenomenological corpus-based approach. More specifically, the research explored what it meant to be a mentor as well as what was involved in experiencing mentoring in a teacher development programme for teachers of English as a foreign language through the analysis of a large corpus of the records of the interaction among the participants in the mentoring. The participants were twenty mentors and three hundred and thirty-five mentees working as English as a Foreign Language teachers in twenty-five branches of a language institute in southern Brazil.

There has been extensive research that have investigated mentoring in various contexts such as business (Carruthers 1993; Mullen 1999; Garvey 2006; Lane 2004; Eby, Rhodes and Allen 2007), science (Jasper 2003; Dean 2009) and education (Scherer 1999; Cullingford 2006; Jones 2006; GEUDER, LANGE and SCAFIDI 2011; Kane and Russell, 2005). In previous research, the concept of mentoring has generally been associated with novice professionals that took part in (instructional) coaching schemes (Clutterbuck and Lane 2006; Grant 2006; Pask and Joy 2007; Law, Ireland and Hussain 2007; Parsloe and Leedham 2009; Nogueira 2011; Cunha 2014) whereby experienced mentors were assigned a novice protégé to guide and counsel in their early years of apprenticeship. In contrast, the present study examines the mentoring process of reasonably experienced teaching professionals instead of focusing on the mentoring process of novice teachers. In addition, in the current study, the focus is on mentees developing their level of awareness and reflection towards their teaching practices and becoming more autonomous professionals.

As such, in this investigation, mentors "help the mentee develop his or her own wisdom [instead of] giving the protégé the benefit of his or her wisdom" (Meggison et al. 2006:17). In the context of this study, mentoring is understood as a developmental process leading to deep professional transformation or changes, instead of being a programme aimed at providing prescriptive models of tutorials aimed at improvement of pre-determined technical competences of novice teachers.
2. Theoretical framework

Four are the theoretical pillars on which this study is based: reflection in/on action (Schön 1983, 1998/2000), the KASA teaching decision-making model (Freeman 1989), the experiential learning model (Kolb 1984), and the model of awareness used to describe communication process in human interactions (Johari Window by Luft 1969). As well as taking part of a 60-hour induction course where they were not only informed of the objectives of the mentoring programme of the institution, but they also took part in input sessions involving topics considered relevant for the mentoring work, such as lesson planning, teaching different proficiency levels, monitoring and giving feedback, dealing with conflict, etc, these mentors also participated actively in 16 meetings held every other fortnight, throughout the year. These meetings served the purpose of building genuine "communities of practice" (Lave and Wenger 1991; Wenger, McDermott and Snyder 2002) amongst the mentors and created a space of mutual trust where they could share their expectations, anxieties, interests, as well as exchange experiences and talk about their needs, worries and frustrations. The mentoring work was based on collaborative cycles of work (Fullan and Hargreaves 2000; John-Steiner 2000) between mentors and mentees which revolved around the planning, the execution, the observation, the reflection and new action plans based on lessons given by the mentees.

3. Method

In order to carry out this research investigation, a mixed-methods approach was adopted that combined both quantitative and qualitative frameworks. The corpus used in this investigation consisted of all of the written communication produced by the participants during 2012 as part of the program. The 979 texts were classified into two types: institutional and spontaneously-generated texts. The institutional texts consist of 920 lesson observation reports written by mentors after the observation of their mentees' lessons. The spontaneous texts were produced by the several different participants in the program on a voluntary basis. These texts ranged from twenty-seven email exchanges between mentors and the researcher, feedback reports written by seventeen school managers on the mentors' performance, feedback reports written by six mentees on their mentors' performance, six case studies written by mentors, to three needs analysis reports written by mentors on two different occasions, in the beginning of the programme and before it ended. The corpus was manually tagged by (i) type of document, (ii) participant, (iii) period of time, and (iv) place. After the manual tagging process of all the texts in the corpus, these texts were then subdivided into smaller units of texts based on the assigned tags through a computer script. A factor analysis was run which enabled us to group the whole corpus into specific factors. The factor analysis yielded five strong dimensions related to the phenomenon of mentoring and that of being a mentor: (i) teacher
roles and functions, (ii) students' perception of progress and learning, (iii) lessons and classroom management, (iv) awareness of changes and (v) teaching procedures. This initial computational analysis was then complemented by a hermeneutic-phenomenological analysis called thematization (van Manen 1990; M. Freire 2010) which involved a manual and systematic analysis of the corpus in search of 'units of meaning' which would lead to the identification of possible themes through which the phenomenon of mentoring and that of being a mentor could be better understood.

The corpus was then submitted to a lexical multi-dimensional analysis (MD; Berber Sardinha, in press), which is a derivation of the functional MD analysis method developed by Biber (1988). The goal of a lexical MD analysis is to detect the correlated groups of lexical items in a corpus, which in turn signal the underlying 'dimensions' of lexical variation. These dimensions in turn indicate the semantic preferences associated with the texts in the corpus. The lexical MD analysis was carried out through a computer program especially designed for this investigation that tagged the corpus for part-of-speech, retained content words (nouns, adjectives, main verbs), normed the counts of each word lemma across the corpus to a rate per 1,000,000 words, then selected the most frequent lemmata in each corpus section (lesson observation reports, email messages, branch managers' reports, mentees' reports, case studies and needs analysis reports). These counts were then entered in an initial factor analysis, the eingenvalues of which were plotted in a scree plot that suggested the existence of five factors. A second, rotated factor analysis (using Promax) was run on the data, and the major loadings (greater than 0.2) were selected. Each factor was interpreted as a dimension of lexical variation, which were used as signposts for the manual identification of the phenomena underlying the mentoring program.

To the best of our knowledge, this is the first use of MD analysis in conjunction in a Hermeneutic-Phenomenological (HP) research design. In a mainstream HP project, the data are fully analyzed manually, but in such studies the data are restricted to a few texts. In this project, however, given the multitude of texts, a full manual analysis would not be reliable, and therefore a computer-based analysis was carried out as a first pass on the data to extract the major lexical configurations, which were then further elaborated on through an extensive hand-and-eye analysis. As a result, the dimensions were reinterpreted into four major HP themes.

4. Results

On the basis of the four HP themes identified, this study revealed that being a mentor as well as what is involved in experiencing mentoring in a teacher development programme can be best understood as being structured around four major themes: transformations, relationships, roles and reflection - which are then subdivided into subthemes. Theme one, transformations, showed that both mentors and their mentees have experienced two types of changes:
improvement of their technical competences and development which led to deep changes in their teaching/mentoring practices; theme two, relationships, involved three subthemes: cooperation, resistance, intervention. Being a mentor meant working in a collaborative mode while being also aware of and prepared to work with resistant professionals, this also lead to more direct interventions from the part of the mentor in changing their mentees’ working practices. Theme three, roles, showed some of the most typical attributes and functions performed by mentors in this mentoring programme, and finally, theme four, reflection, confirmed that in any practice aiming at changes, it is a sine qua non condition to have professionals with a good level of awareness and critical reflection on their practices.

References


‘Do animals have nationality?’ Insights from a cross-linguistic corpus-assisted study on animal representation in online newspapers
Anda Drasovean (King’s College London, UK)

In recent years, animals have become an increasingly visible topic of research in social sciences and the humanities. Much of this work stems from the idea that, in a post-industrialist society in which real-life animals are “disappearing” (Berger, 2009: 11), looking at representations of animals is essential to understanding their role in human society (Baker, 2001: 190). More importantly, these representations are historically and culturally variable (e.g. DeMello, 2012; N. Taylor, 2013). As such, they not only have the potential to reflect how animals are conceptualised, classified, and regarded by different cultures and discourse communities, but also to reveal some of the peculiarities of these cultural contexts.

So far, the linguistic studies on this topic have limited their analyses of animal representation to a single language (e.g. Jepson, 2008; Sealey & Charles, 2013; Stibbe, 2003; inter alia). One exception would be the small body of studies that employ the cognitive metaphor framework to explore how conventional animal metaphors in different languages might serve as a way of exploring cultural differences (e.g. Hsieh, 2006; Talebinejad & Dastjerdi, 2005). However, these studies are, by definition, limited to the analysis of figurative language – that is, to how animal metaphors are used to talk about humans, rather than animals.

By exploring animal representation from an empirical, quantitative, and cross-linguistic perspective, the study presented in this paper aims to fill a methodological gap in the linguistics literature on this topic. The data consists of two comparable, topic-based purpose-built corpora of contemporary (2012-2014) online news stories about animals – one in Romanian, and one in English. Each corpus is organised in five subcorpora, based on the five subtopics/case-study animals used to collect it: dogs, cats, bears, pigs, and horses (these were found to be the most frequently mentioned animals in a general corpus of Romanian and one of English). To analyse this data, I have employed a methodological approach known as Cross-linguistic Corpus-Assisted Discourse Studies (C-CADS) (Partington, Duguid, & C. Taylor, 2013; Vessey, 2013), which was extended and tailored for the specific requirements of the present study. In the initial stage of the analysis, I used keyword analysis and topic modelling to identify the most salient themes in each subcorpus. I then explored these salient themes further with the help of bilingual word sketches (Kilgarriff & Tugwell, 2002), and collocation and concordance analysis.

A key finding emerging from the analysis, on which I focus in this paper, is that animals are routinely used in online newspaper discourse as a means of highlighting and discussing cultural characteristics and differences. To explore this issue in relation to Romania and the UK, I started by looking
at the frequencies of country names and national adjectives and their distribution across the two corpora of animal news stories. The normalised frequency figures show that Britain is approximately four times more frequently mentioned in Romanian news than Romania is in British news. This might be explained by the fact that Britain, as a prestige nation, has high newsworthiness for Romanian media; additionally, the global outreach of British newspapers such as MailOnline seem to make these a prolific news source for Romanian tabloids. Romania is mostly mentioned in the British online press in relation to some very particular situations, such as the horsemeat scandal or Romania’s stray dog issue. Based on the distribution of these country names and national adjectives across the files in each subcorpus, I was able to identify some typical uses of these terms. I explored these further using concordance analysis and by qualitatively analysing a small sample of prototypical texts.

Two patterns of use seemed particularly salient: (1) the use of animals or animal products as symbols of national identity; and (2) using animals as proxies for representing other cultures and/or discussing cultural differences. In what concerns the former, the Romanian corpus offers an interesting example of how pigs – or, more specifically, the slaughtering of pigs before Christmas – are framed in the online press as a symbol of Romanian traditions. Despite the controversy generated by such practices, which do not abide by EU animal welfare standards, the corpus data suggests that the traditional slaughtering of pigs is described in Romanian news stories as ceremonial and picturesque, and an important part of Romanian identity. By contrast, the phrase a nation of animal lovers, which occurs most frequently (38 occurrences) in the subcorpus of UK news about dogs, suggests that British online news stories reinforce the idea that love for animals is an important part of British identity. Moreover, the idea that Britain is a nation of animal lovers, which seems to implicitly involve consideration for animal welfare, seems to often serve as an argument in negatively stereotyping other nations based on their alleged mistreatment of animals.

Thus, the second pattern of use, which is typical of the British corpus, entails the negative portrayal of Romania and Romanians in animal news stories. Concordance analysis for the lemmas ROMANIA and ROMANIAN seems to corroborate the findings of Ibrahim and Howarth (2016), who showed that, in their coverage of the horsemeat scandal, British newspapers have negatively portrayed Romanian and Eastern-European migrants. Furthermore, the analysis shows that Romania’s problem with stray dogs was also used by a part of British media to promote an anti-immigration discourse. As can be observed Fig. 1, the articles about stray dogs go beyond a vivid and generalising description of Romania’s cruelty towards dogs, to questioning Romania’s membership in the EU, or suggesting that Romanian immigrants are dangerous and dishonest:
he had served time in jail for theft and dishonesty in Romania and Italy. He was also handed a conditional

this of dead dogs lying in skips are far too common in Romania [PH] “Funds are provided to humanely euthanize

to Britain from the EU’s two poorest countries - Romania and Bulgaria - lapsed. - Intimidation and violence

with new legislation encompassing animal rights. If Romanians truly want to join Europe, they need to see the

Figure 1 – Concordance lines for ROMANIA\ROMANIAN

Amongst concerns about increased immigration from Romania and Bulgaria, a part of the British online press seems to have framed animal related topics such as the horsemeat scandal or stray dogs as a means of indirectly expressing xenophobic attitudes.

An interesting question emerging from these findings concerns the extent to which national adjectives can be used with animals. In the two corpora, the adjectives românesc (‘Romanian’) and British are in general used in two different contexts: in relation to national economic activities and interests (e.g. British farmers); and in relation to traditional foods or products (e.g. British bacon). Much less frequently, phrases such as Romanian dogs or Romanian refugee dogs occur in the British corpus. In the light of the previously discussed findings, such atypical formulations can be interpreted as a means of intertwining emotional stories about abused dogs with an anti-immigration discourse.

Overall, the proposed paper highlights a less discussed aspect in the linguistic literature about animal representation. In addition to revealing a range of cultural and linguistic differences and similarities between Romania and the UK, the comparative analysis of two corpora of animal news stories also suggests that animals are actively and habitually used in online newspaper discourses as surrogates for specific nations or ethnic groups. In my presentation, I will discuss several such situations, and illustrate them with examples from concordance lines and representative texts.

References


The pragmatics of insubordinate if -clauses in British English: A corpus-based study
Cristina Lastres-López (University of Santiago de Compostela, Spain)

Conditionals are prototypically formed by two clauses: a subordinate clause, most frequently introduced by the conjunction if in English, known as protasis (p); and a main clause, known as apodosis (q), as illustrated in (1) below.

(1) If it's a really nice day, we could walk (ICE-GB: S1A-006 #301: 1: B)
   \[ p \rightarrow q \]

However, clauses introduced by if can also stand in isolation in discourse, without the presence of an apodosis, as in (2).

(2) Uhm <,,> perhaps if you could tell me a little bit about your own father <,,> (ICE-GB: S1A-072 #042: 1: A)

While full conditionals as (1) have been widely studied (Traugott, Meulen, Reilly, & Ferguson, 1986; Athanasiadou & Dirven, 1997; Declerck & Reed, 2001; Dancygier & Sweetser, 2005; to name but a few), if -clauses such as the one presented in (2) have traditionally been neglected or treated as marginal cases in grammars (Quirk, Greenbaum, Leech, & Svartvik, 1985: 841-842; Huddleston & Pullum, 2002: 945). Evans (2007: 367) coined the term insubordination to refer to this grammatical phenomenon, which can be defined as “the conventionalised main clause use of what on prima facie grounds appear to be formally subordinate clauses”. In these constructions, the apodosis is not required and neither is it considered to be a case of ellipsis, since the clause no longer expresses a conditional relation between p and q, as in (1) above. Instead, example (2) issues a directive, thus being an alternative, pragmatically, to either an interrogative, as shown in (3a) or an imperative, as in (3b).

(3a) Could you tell me a little bit about your own father?
(3b) Tell me a bit about your own father

This paper aims to complement prior studies that have examined insubordination across different languages and have reported that these constructions may express a varied range of functions (Schwenter, 1996, 2016; Stirling, 1999; Evans, 2007; D’Hertefelt, 2015; Sansiñena, De Smet, & Cornillie, 2015; Evans & Watanabe, 2016; Kaltenböck, 2016). In line with Kaltenböck (2016), this study focuses on British English and resorts to corpus rather than elicited data. The aim of this paper is to contribute to throw light on the diversity of functions these constructions may have in discourse, analysing insubordinate if -constructions in British English and elucidating their uses and functions in spoken discourse. The analysis also focuses on the frequency of these constructions and on alternative grammatical patterns that may also be possible to express the same discourse function, as in (3a) and (3b) above. Furthermore, the study tests whether certain types of verbs increase the
likelihood that the clause expresses a certain function, and can therefore be considered predictors of the function of the clause. For this purpose, verbs are considered both in their classification modals/non-modals and in the semantic taxonomy proposed by Halliday and Matthiessen (2014).

The methodology used is corpus-based, analysing manually data extracted from two different sources: the British component of the International Corpus of English (ICE-GB) (Nelson, Wallis, & Aarts, 2002), a one-million word corpus; and the British National Corpus (BNC), one hundred times bigger in size. On the one hand, the results obtained from ICE-GB offer an overview of the frequency and use of these constructions across a variety of spoken genres and reveal that they are more conspicuous in conversation, with more than 60% of the tokens extracted from this genre. Due to the infrequent nature of the constructions under analysis, the results obtained from ICE-GB require to be further complemented by a larger sample. Thus, on the other hand, the data from the BNC concentrate on the type of text in which they appear more recurrently. The subcorpus of conversations of the BNC, twenty times bigger than its counterpart in ICE-GB, is selected to examine and illustrate different uses and functions of if -insubordination in conversation. The results obtained from the analysis display that the constructions under examination may express a wide range of functions in discourse. Among them, the results report that insubordinate if -clauses most frequently express polite requests, but that up to seven other functions are attested, with wishes and suggestions, as in (4) and (5) occupying the second and third most frequent functions for these constructions in discourse.

(4) If we had a regular bus service [pause] rather than a regular train service (BNC, KCS 332)
(5) A: so you can take them as well, don’t they? (BNC, KC2 1146)
   B: Mm, mm
   A: If you book serve
   B: Serve, yeah Serve takes them, yeah I know

References


The Fragment Corpus (FraC)
Eva Horch and Ingo Reich
(Universität des Saarlandes, Germany)

We present the Fragment Corpus (FraC), a corpus for the investigation of fragments (see Morgan 1973), i.e. incomplete sentences, in German. The corpus is a mixed-register corpus and consists of 17 different text types including written (newspaper texts, legal texts, blogs etc.) and spoken texts (dialogues, interviews, radio moderations etc.), as well as social media texts (tweets, sms, chats). Each subcorpus comprises approx. 2000 utterance units (including fragments, following the orthographic notion of sentence) which amounts to a total corpus size of 380K tokens. The data was taken from electronically available sources.

Depending on availability, metadata include information on source, author, date and text type. The corpus is annotated for tokens, lemmas and POS using TreeTagger (Schmid 1994, 1995). In addition, each utterance unit is manually annotated with:

- its syntactic category (like VP, NP or SEQUENCE for sequences of syntactic phrases)
- a listing of immediate constituents (in the case of sequences or coordinations)
- the discourse function of the utterance unit (like greeting, listing, discourse topic)
- idiosyncratic information (like ‘containing an acronym’, ’formulac expression’)
- omission type (like topic drop, object drop, article and copula omission)
- null elements (for not realized elements in case of article and copula omission)

FraC was built within the project B3 of the CRC 1102 on Information Density and Linguistic Encoding (IDeaL). In this project, the corpus is primarily used for corpus linguistic studies of fragments as well as for building language models in order to investigate the actual use of fragments by relating their use to information-theoretic notions like ID (see Shannon 1948). It is planned to also annotate the corpus with Surprisal values for each token based on trigrams and to parse the corpus (with the Stanford parser, Rafferty & Manning 2008).

The source of the corpus is in XML format. The corpus is encoded in two different formats: (i) in CQP format (cf. IMS Open Corpus Workbench CWB; Evert & Hardie, 2011) for query and statistical analysis and (ii) in ANNIS format (Krause & Zeldes 2016). The corpus will be made accessible through the ANNIS multilayer corpus tool and the CQPweb installation of Saarland University under the current copyright laws. In the medium term, we would like to make FraC available through the CLARIN-D repository of Saarland University.

References


Conference on New Methods in Language Processing, 44–49. Manchester, UK.
In Proceedings of the ACL SIGDAT-Workshop, Dublin, Ireland.
The investigation of Dickens’s style using corpus tools is not new (see, among others, Hori, 2004; Mahlberg and Smith, 2012; Mahlberg, Smith and Preston, 2013; Mahlberg, 2013; Stockwell and Mahlberg, 2015; Ruano San Segundo, 2016). This presentation follows in their footsteps. Specifically, I look into Charles Dickens’s use of direct thought presentation in his fifteen novels (c. 3.8 million words) using a corpus-stylistic approach. The full texts were downloaded from the Project Gutenberg website and were processed using the WordSmith Tools 7 (Scott, 2016) software. The aim of the analysis is to delve deeper into Dickens’s presentation of his characters’ thoughts, an aspect so far underexplored maybe due to the ‘lack of psychological inwardness and depth in his characters’ (McParland, 2011: 209). Despite such dearth of psychological depth, though, Dickens consistently reported his characters’ thoughts throughout his fifteen novels. Therefore, a systematic analysis of how he did so seems in order, if only because no comprehensive account of it has been yet attempted. As will be shown, occurrences of direct thought (henceforth, DT) can be effectively retrieved thanks to a corpus methodology, which makes it possible to systematically analyse Dickens’s use of this mode of thought presentation. The analysis will focus on those examples that contain the verb think, the reporting verb for thought presentation par excellence. For example:

(1) ‘John’ thought madame, checking off her work as her fingers knitted, and her eyes looked at the stranger. ‘Stay long enough, and I shall knit ‘BARSAD’ before you go.’ (A Tale of Two Cities, book 2, chapter 16)

This example contains several characteristic features of Dickens’s use of DT, such as the use of a vocative in the reported clause, a suspended reporting clause and the reference to the character’s eyes. These and other traits will be investigated in this presentation. As will be shown, they fulfil meaningful functions which relate to significant aspects of Dickens’s style, as discussed by other critics. As far as the search for instances of think is concerned, several concordance searches were made following the patterns under which the verb is used: ‘thought, ”thought, ’thinks,”thinks,’he thought,’he thinks,”he thinks,’she thought,”she thinks,’I thought,”I thought,’I think and ”I think. Figure 1, for instance, shows twenty-five examples of the 189 hits retrieved from Dickens’s fifteen novels after the concordance search ‘thought.’

---

1 As Project Gutenberg texts present irregularities when it comes to the use of inverted commas — texts do not necessarily stick to either simple or double inverted commas consistently —, every concordance had to be made twice.

2 This concordance search identifies all the occurrences of DT in which, as is the case of the example from A Tale of Two Cities shown above, the reporting verb (thought in this case) is used immediately after the inverted comma closing the character’s thoughts.
The concordance searches retrieve a total of 244 occurrences of *think* as a DT reporting verb in Dickens’s fifteen novels, constituting a much wider set of examples than the twenty-one examined by Busse (2010) in the most comprehensive analysis of discourse presentation strategies in nineteenth-century fiction to date.3

The analysis of these 244 occurrences will provide solid evidence in support of previous findings on thought presentation in nineteenth-century English narrative fiction, such as the relationship between characters’ glances and the presentation of their thoughts or the use of DT in moments of heightened intensity. However, this analysis will also unveil hitherto unremarked patterns in form and function as far as Dickens’s presentation of his characters’ thoughts is concerned: the relationship between body language and the presentation of characters’ thoughts (the fact that women are presented in a sitting position, for example), the character information presented in reporting clauses, or the use of suspensions to create an effect of synchronicity between body language and thought presentation, among others. The analysis, in sum, is intended to contribute to a better understanding of Dickens’s craftsmanship using corpus tools.

The presentation will be divided as follows. I will begin with a brief overview of DT as a strategy for reporting mental discourse and its uses and functions in nineteenth-century English fiction. Next, the methodological procedure used to retrieve the examples will be explained and the results obtained will be presented. The analysis of these results will be divided into four

---

3 It is only fair to note that Busse’s corpus is composed of excerpts of less than 3,500 words from twenty-two nineteenth-century novels (Busse 2010: 64), being therefore much smaller than the corpus of Dickens’s novels analysed here.
parts. Firstly, I will discuss the overall distribution of DT across the corpus. Secondly, I will discuss the most striking aspects of Dickens’s use of DT from a formal perspective (the use of parentheses enclosing instances of DT, vocatives and exclamations). Thirdly, I will analyse Dickens’s use of reporting clauses in DT, focusing on aspects of body language and characterisation. Finally, I will also examine the use of reporting clauses as suspensions; this will allow me to demonstrate that Dickens used these in a very much similar manner to those in direct speech. To do so, the CLiC tool (Corpus Linguistics in Cheshire; see http://clic.bham.ac.uk/) was used, which allows for the systematic retrieval of suspensions from literary texts.

References


Combining corpus and experimental methods to investigate trust-building strategies in corporate discourse
Matteo Fuoli (University of Birmingham, UK) and Christopher Hart (Lancaster University, UK)

Introduction

This study combines corpus-based and experimental methods with the aim to (i) identify and (ii) test the effects of several key trust-building strategies used by companies in their public discourse. First, a corpus-based analysis of a specialized corpus of annual and corporate social responsibility (CSR) reports is conducted in order to investigate how companies discursively construct a positive corporate identity in these texts in an attempt to gain the readers’ trust. A scenario-based experiment is then carried out to empirically test the persuasiveness of some of the most salient trust-building strategies identified in the corpus analysis. Below we provide details about the two parts of the study, in turn.

The corpus-based analysis

The corpus-based analysis examines how companies use stance expressions (e.g. Biber and Finegan, 1989; Biber et al., 1999; Conrad and Biber, 2000) in annual and CSR reports to construct and promote a positive corporate identity and gain the trust of the stakeholder groups that these texts target. The analysis addresses the following research questions:

Q1: Are there differences in the frequency and type of stance expressions used in annual and CSR reports?
Q2: How are stance expressions used to discursively construct and promote a trustworthy corporate identity in these two text types?
Q3: Do companies present themselves differently in annual and CSR reports? And, if so, how do they discursively shape their identity to appeal to and earn the trust of the audiences that these texts target?

The analysis is based on specialized corpus of roughly 2.5 million words, which includes the 2011 annual reports and CSR reports published by a sample of publicly traded multinational corporations belonging to four different industry sectors. Table 1 provides summary information about the corpus.

The analysis is based on a slightly adapted version of the framework for the analysis of stance presented in Biber (2006b: Chapter 5). The framework comprises three main categories of stance devices: attitudinal stance, epistemic stance and modality. Markers of style-of-speaking stance were not considered, as they proved to be extremely infrequent in the corpus. As in Biber (2006b), modality was kept separate from epistemic stance, even though modals and semi-modals may convey epistemic meanings when used in the extrinsic sense, and could therefore, in principle, be assimilated to the category of epistemic stance.
Table 1: Overview of the corpus

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of reports</td>
<td>32 (16 annual reports; 16 CSR reports)</td>
</tr>
<tr>
<td>Total corpus size</td>
<td>2,433,956</td>
</tr>
<tr>
<td>Size of subcorpora</td>
<td></td>
</tr>
<tr>
<td>ACR</td>
<td>1,828,029</td>
</tr>
<tr>
<td>CSR</td>
<td>605,927</td>
</tr>
<tr>
<td>Mean report size</td>
<td></td>
</tr>
<tr>
<td>ACR</td>
<td>114,251.8</td>
</tr>
<tr>
<td>CSR</td>
<td>37,870.4</td>
</tr>
<tr>
<td>Sectors and companies</td>
<td></td>
</tr>
<tr>
<td>Financial services</td>
<td></td>
</tr>
<tr>
<td>JP Morgan Chase</td>
<td></td>
</tr>
<tr>
<td>HSBC</td>
<td></td>
</tr>
<tr>
<td>Wells Fargo</td>
<td></td>
</tr>
<tr>
<td>Bank of America</td>
<td></td>
</tr>
<tr>
<td>Oil and gas</td>
<td></td>
</tr>
<tr>
<td>Exxon Mobil</td>
<td></td>
</tr>
<tr>
<td>Royal Dutch Shell</td>
<td></td>
</tr>
<tr>
<td>Chevron</td>
<td></td>
</tr>
<tr>
<td>BP</td>
<td></td>
</tr>
<tr>
<td>Pharmaceuticals</td>
<td></td>
</tr>
<tr>
<td>Johnson &amp; Johnson</td>
<td></td>
</tr>
<tr>
<td>Merck</td>
<td></td>
</tr>
<tr>
<td>GlaxoSmithKline</td>
<td></td>
</tr>
<tr>
<td>Abbott Laboratories</td>
<td></td>
</tr>
<tr>
<td>Food processing</td>
<td></td>
</tr>
<tr>
<td>Neslté</td>
<td></td>
</tr>
<tr>
<td>Unilever</td>
<td></td>
</tr>
<tr>
<td>General Mills</td>
<td></td>
</tr>
<tr>
<td>Kellogg</td>
<td></td>
</tr>
</tbody>
</table>

The procedure for the identification and quantification of stance constructions in the corpus relied on a combination of automatic techniques and manual analysis. First, an extended list of stance markers was created based on the lists provided in Biber and Finegan (1989) and Biber (2006b). By means of a string search Perl script, the stance markers included in the lists were simultaneously and automatically searched for in the corpus. Second, the results of the automatic search were manually inspected in order to identify and remove all false positives. This was accomplished by means of concordance analysis with the aid of the corpus program AntConc (Anthony, 2012). Once concordances for all the stance markers retrieved in step one were produced and all irrelevant entries were removed from them, the frequency of occurrence of each marker in the annual report and CSR report subcorpora was calculated. Frequencies were normalized to allow for a comparison between the subcorpora, which are of different size. Where relevant, pairwise log-likelihood tests (Rayson and Garside, 2000) were performed to determine whether the differences observed in the frequency of stance markers between the annual report and CSR report subcorpora were statistically significant.

The results of the corpus-based analysis show that companies profile distinct identities in annual and CSR reports. In annual reports, they use stance resources to portray themselves as unbiased, rational, and competent decision makers. In CSR reports, they present themselves as committed, honest, and caring corporate citizens. These discursive identities are interpreted as strategic self-representations that optimize the

---

1 The script, which is called ‘filter_by_field.pl’, is included in the supplementary materials for the following article: M. Baroni and A. Lenci. Distributional Memory: A general framework for corpus-based semantics. 2010. Computational Linguistics 36 (4): 673-721. It can be downloaded, free of charge, from this URL: http://clic.cimec.unitn.it/dm/materials/scripts.tar.gz [Last accessed: 20 May 2016]. Instructions on how to run the script are available on demand.
persuasive appeal of the reports by addressing the specific expectations of the target readerships.

The experiment

Similar to previous analyses of corporate texts (e.g. Bondi, 2016; Breeze, 2013; Hyland, 1998; Rutherford, 2005), the corpus-based analysis is limited in that it does not provide empirical evidence of the actual persuasiveness of the discourse strategies it brings to light. This limitation is, indeed, common to all qualitative as well as corpus-based discourse analysis research. The experimental part of this study provides the next logical step towards assessing the persuasiveness of corporate narratives. As a follow-up to the corpus-based analysis, we developed and conducted a scenario-based experiment designed to measure the effectiveness of some of the key trust-building discourse strategies that, as the analysis has revealed, are commonly employed by companies in their annual and CSR reports.

The study incorporated a 2x2 between-subjects design. Participants were asked to read a fabricated news story about a fictitious pharmaceutical company who is accused of having paid illegal kickbacks to doctors to help persuade them to prescribe their products over cheaper alternatives. The article reported that the company denied the accusations. Two versions of the article were presented, one in which evidence of the company’s guilt was presented as weak, and one in which it was presented as strong. Before reading the article, half of the participants (i.e. the test group) read a fabricated ‘About Us’ page from the accused company’s website. The webpage contained some of the discursive strategies that, based on the corpus-based analysis, we hypothesize perform a trust-building function. Upon reading the text(s), participants completed a questionnaire designed to measure their trust in the accused company.

The results of a pilot study involving 44 individuals show that the discursive strategies contained in the ‘About Us’ page had a positive effect on participants’ trust in the accused company, regardless of the strength of the evidence. As shown in Fig. 1, the participants who were exposed to the ‘About Us’ text had more positive impressions of the accused company’s integrity and benevolence, and perceived the company’s denial as more credible compared to the participants who were not exposed to the text. Multiple regression analyses reveal that the differences between the groups are statistically significant.

The results of this study, which are shown in Figure 1 below, provide novel insights into the discursive dynamics of organizational trust, and demonstrate the usefulness of experimental methods as a means for triangulating the findings of qualitative and corpus-based discourse analysis research (Hart, in press).
Figure 1. Variable means by experimental condition

<table>
<thead>
<tr>
<th></th>
<th>Perceived benevolence</th>
<th>Perceived integrity</th>
<th>Perceived credibility of denial</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant ratings</td>
<td>Weak</td>
<td>Strong</td>
<td>Weak</td>
</tr>
<tr>
<td>Perceived benevolence</td>
<td>3.00</td>
<td>2.00</td>
<td>3.00</td>
</tr>
<tr>
<td>Perceived integrity</td>
<td>3.00</td>
<td>2.00</td>
<td>3.00</td>
</tr>
<tr>
<td>Perceived credibility of denial</td>
<td>3.00</td>
<td>2.00</td>
<td>3.00</td>
</tr>
</tbody>
</table>

Below are three interaction plots, one for each dependent variable, showing the observed mean scores. As we can clearly see from these plots, the 'about us' page has a pretty strong effect on the participants, in all of the dimensions considered. The strongest effect was on perceived integrity. This is really encouraging, I think.

### Perceived benevolence

- Weak: 3.00
- Strong: 2.00

### Perceived integrity

- Weak: 3.00
- Strong: 2.00

### Perceived credibility of denial

- Weak: 3.00
- Strong: 2.00
References


In November 1864, only 5 months before he was assassinated, Abraham Lincoln, the 16th President of the United States, sent a short letter of condolence to the widow Lydia Bixby of Boston, the mother of five sons who were believed to have died in the Civil War. In fact, the Widow Bixby had only lost two sons and was also a Confederate sympathiser, who destroyed the letter in anger after reading it. Fortunately, the Adjutant General of Massachusetts, William Shouler, who had requested the letter from Lincoln in the first place, sent a copy to the Boston Evening Transcript, where the letter was published the following day. The Bixby Letter would go on to become one of America’s most renowned pieces of correspondence and one of Lincoln’s most celebrated texts, surpassed in notoriety only by the Gettysburg Address and the Emancipation Proclamation. It has also become part of public culture, for example being prominently featured in the movie Saving Private Ryan and being recited by President George W. Bush at the 10 year anniversary of the September 11th Attacks.

Despite its fame, the authorship of the letter has long been in dispute, with some historians claiming, based primarily on external evidence, that Lincoln’s young personal assistant, John Hay, the future Secretary of State under McKinley and Roosevelt, was its author. Most notably, Burlingame (1995, 1999) argued that Hay was the author of the letter based on historical and linguistic evidence, including the use of the word beguile, which only Hay is known to have used in his writings. Overall, however, the debate is still unresolved and linguistic evidence has been minimal and inconclusive. The goal of this paper is therefore to investigate whether Lincoln or Hay is the more likely author of the Bixby Letter based on a quantitative comparison of the style of the letter to relatively large corpora consisting of the known writings of both statesmen.

First, all accessible writings of both Lincoln and Hay, which are voluminous especially for Lincoln, were obtained online from various historical archives. The texts were checked by hand to identify editorial notes and changes and to exclude co-authored and other problematic documents. We also choose to exclude all Lincoln’s texts that were written after 18 May 1860, when Lincoln was nominated as the presidential candidate of the Republican Party, because Hay became Lincoln’s personal secretary soon after that date. In total, the Lincoln contains corpus 1,085 texts totalling approximately 400,000 words, including substantial numbers of letters, speeches, bills, and resolutions. The Hay corpus contains 577 texts totalling approximately 260,000 words, including substantial numbers of letters, prose, poems, and diary entries.

This case of disputed authorship is especially challenging because the Bixby letter is so short. Totalling only 139 words, the Bixby letter is far too short to be attributed using standard stylometric techniques for authorship attribution, which generally requires disputed texts to be at least 500-1,000 words long so that the
relative frequencies of numerous linguistic features (e.g. function words) can be estimated accurately (Grieve, 2007). The usual approach to analysing short texts, for example in forensic linguistics where short texts are the norm, is therefore to look at whether or not forms in the questioned document occur in each possible author writing sample, ideally showing that the vast majority of linguistic forms found in the questioned text are only used by one possible author. There are, however, at least two major issues with this approach: how to select an unbiased feature set and how to control for variation in sample size.

Based on this general approach, and keeping these two limitations in mind, we have developed a new quantitative method for attributing short texts, which we refer to as n-gram tracing. The basic idea behind the method is to calculate the percentage of all n-grams in the questioned document that occur in each of the possible author writing samples. An n-gram is a sequence of one or more (e.g. 1-gram, 2-gram, 3-gram, etc.) linguistic forms in a text, which can be measured at any level (e.g. character-level, word-level). To conduct n-gram tracing, first all n-grams of a particular length and level (e.g. word-level 2-grams) are extracted from the questioned document. The percentage of those n-grams that occur in each of the possible author writing samples is then calculated. To control for variation in sample size, the percentages of forms are calculated for random samples of texts of different sizes drawn from each possible author writing. The possible author that uses a higher percentage of n-grams, especially as the size of these sample increases, is then selected as the most likely author of the disputed text.

To evaluate the method, we tested it using 1-4 word-level n-grams and 1-20 character-level n-grams on all 1,662 texts of known Lincoln and Hay authorship in our corpus. Specifically, we would remove one known text from our corpus, extract all the n-grams from that text, and then trace those n-grams across the remaining Hay and Lincoln texts to attribute the text. We have found the method to be highly accurate. Most notably, when attributing texts based on 4-10 character-level n-grams, the method attributes all 1,662 texts in the corpus of possible authors correctly. Remarkably, a majority of these texts contain fewer than 200 words and 10% contain fewer than 50.

Finally, to attribute the Bixby Letter we extracted all 1-4 word and 1-20 character level n-grams and then traced each set of n-grams across the Hay and Lincoln corpora. All analyses clearly identified Hay as the most likely author of the Bixby Letter. We therefore conclude that Hay as opposed to Lincoln was the author of the Bixby Letter, providing linguistic support for the analyses of Burlingame and others. N-gram tracing also appears to offer a solution to the problem of text length in authorship attribution, one of the most important problems in stylometry.

References


The present research focused on the possible differences in the usage of hedging strategies in linguistic research articles of native Czech speakers writing in English (NCS) and native speakers of English (NES). Hedging is considered to be a fundamental part of any academic text ensuring that the outcomes of one’s research are successfully communicated to the intended audience. The hypothesis that there will be differences in the use of hedging strategies was based on previous work done by Čmejrková et al. (1999) and Dontcheva-Navrátilová (2015) who have shown significant differences between the two groups in the conventions and approach to presentation of one’s research to the academia. Czech academic writing is characterized as being less interactive, primarily writer-oriented, and concealing authorial presence in comparison to Anglophone academic writing, which exhibits higher level of interactivity, is primarily reader-oriented, and often shows marked authorial presence.

The research articles providing data for this study were collected from international and Czech based peer-reviewed journals. Subsequently, two corpora were created, each consisting of 23 concluding sections of research articles and each containing approximately 11,000 words. The data were analysed using the taxonomy of hedges established by Hyland (1996, 1998), employed in Varttala’s work (2001), and further modified by Malášková (2015) who recognizes three hedging strategies – content-oriented (attenuation of claims), writer-oriented (limiting personal commitment of the author), and reader-oriented hedging (subjectivizing claims). The analysis of the data was twofold; firstly the individual lexical items functioning as hedges either by themselves or contributing to a more complex hedging structures were identified and classified. Secondly, the items were categorized according to which hedging strategy they represented.

The results were normalized and compared through log-likelihood to be able to determine significantly different values. The results have shown that the existing conventions in Czech academic writing do influence the use of hedges in research articles of NCS. The most salient differences were found in the category of the reader-oriented hedging strategies (personal reference to methods, personal attribution, inclusive we, questions, etc.) which were present with significantly higher frequency in the corpus of NES. The reader-oriented hedging contributes to higher interactivity and dialogic nature of the text and the results thus comply with the previous research. The category of writer-oriented hedges shows significant difference in the frequency of use of passive constructions (higher frequency in the NCS corpus), a strategy used to avoid overt authorial presence, which is again in compliance with the preliminary hypothesis. Content-oriented hedging strategies were employed consistently and with no significant differences in frequency by both groups, except for the category of adverbs, whose frequency was significantly higher in the NCS corpus. Finally, the overall number of all hedging devices and strategies was higher in the NES corpus. The results of this small-scale study may have implications for teaching academic writing to native Czech authors in order to
achieve academic discourse that is able to better compete in the environment of the Anglophone academic community.

References


Design of test-making tools for the learner corpus

Olga Vinogradova and Ekaterina Gerasimenko (National Research University Higher School of Economics, Russia)

Russian Error-Annotated Learner English Corpus (further REALEC) is an electronic collection of texts written in English by Russian university students studying English at an intermediate – upper-intermediate level. The corpus is easily available at http://realec.org/#/realec/ and includes over 1200 texts (about 360 thousand word tokens), mostly essays written in preparation for IELTS-type English examination taken at the end of the second-year of Bachelor studies, and the newer, growing sub-corpus of 2nd-year Bachelor students’ essays written in the English examination at http://realec.org/#/hse/data_4_staff/IELTS/ consisting by the beginning of 2017 of about 3000 essays comprising almost 700 thousand word tokens. The value of a learner corpus is in its annotation practice [Granger, S. (2003), Leech, G. (2015)], so experts at the Higher School of Economics (teachers and students familiar with annotation approaches adopted in REALEC) annotate errors in the essays with the help of the error classification scheme established at the start of setting up the corpus [Kutuzov, A. and Kuzmenko, E. (2014)]. The error annotation scheme consists of 152 categories organized into a tree-like structure. Annotators are instructed to choose a specific tag for the error they have spotted, and unified approaches to annotation have been worked out [Kutuzov, A., Kuzmenko, E., and Vinogradova, O. (2015)].

REALEC is not only a ground for erratological statistics and analyses with results applicable in language acquisition [Vinogradova, O. (2016)], but also a great resource of sentences that can be used in tests for English learners, as annotated sentences in the corpus have all the necessary elements for generating testing questions fully or semi-automatically, namely, error spans are outlined, errors are labeled, and the corrected option is offered by the annotator. The current paper presents RETM – REALEC English Test Maker, the system that works as a tool to automatically generate tests for students on the basis of the errors that experts have marked in student works submitted to REALEC. With the help of the scripts written in Python (source code can be found at https://github.com/ElizavetaKuzmenko/realec-exercises). RETM extracts the necessary testing questions from Brat and transforms them into the XML file required for the format of the user interface.

At first the script parses the annotation files (initially stored separately for each text) and marks annotations in the texts using different tags for relevant error types and for errors that have to be corrected. Afterwards, the script takes annotated texts, corrects all mistakes and creates questions for relevant mistakes if there are any. The annotations marked in the texts contain the right answer (suggested by the expert annotator in the corpus) and indices that are used to insert words in an open cloze or to

---

1 This article is an output of a research project implemented as part of the Basic Research Program at the National Research University Higher School of Economics (HSE).
highlight the mistake in the question text. Subsequently, questions and right answers are passed to the XML formatter and then uploaded manually to the user interface.

For the sake of the convenience of working with the tests, the interface was set up in Moodle (Modular Object-Oriented Dynamic Learning Environment), the open source learning platform which has progressively been found well applicable in teaching processes (https://moodle.org/). This platform provides a user-friendly environment for instructors to prepare and administer the test they want and get the statistics on their students’ performance after the test. The convenience works equally well for testees as throughout the test they get as much feedback and support as their instructor chooses for them. The tests that have already been applied can be seen at http://web-corpora.net/realec, the space where both instructors and their students undergo the necessary authorization.

After the first introduction of RETM in 2016 [Kustova, M. (2016), Vinogradova, O. (2016-2)], the group of linguists updated the system and expanded its functions by adding new types of tests and giving some new components of the automated generation test-making process a try. At the moment there are about 30 different tests with about 1200 testing questions altogether. These recent achievements are in the focus of the present talk.

The experimental ground for the research was formed with two tests in Moodle. The questions for these tests were selected from the pool of sentences generated at the first stage from the sentences of student essays by correcting all error spans except for one for the option given as the correction by the expert annotator. At the second stage the testing instructor looks at each testing question in order to make sure that:
- there is no more than one error left in the sentence;
- the error span can be clearly identified;
- the context of one, two, or three sentences will be enough for a testee to be able to spot and correct the error.

At this stage a test-maker can edit a question, delete a part of the question, delete a sentence from the pool altogether, and make a decision what kind of test will best fit the testing purposes for this error. In the reported research we have applied the following two test types, with two subtypes in either of them, namely:
- correcting errors (short answer): spot and correct the error, or correct the highlighted error.
- filling gaps (open cloze): give the appropriate form of a word in the parentheses, or fill the gap with the appropriate word or words.

Thus, the overall process of creating and testing questions is the following:

1) extracting sentences containing the error. The annotations are visualised in the user interface:

![Example of visualised annotations](image_url)
In this example, the question is created from the error marked with the tag ‘Choice of tense’.

2) creating questions automatically by processing the relevant error and correcting all other errors. There are two formats of output – XML (for uploading questions to Moodle) and plain text.

   Nowadays, the problem of public health {1:SHORTANSWER:=has become} incredibly important worldwide. Many people all over the world forget about their physical development and health because of a whole new set of opportunities we have today.

3) editing manually:

   Nowadays, the problem of public health {1:SHORTANSWER:="has become ~="}is becoming) (become) incredibly important worldwide. Many people all over the world forget about their physical development and health because of a whole new set of leisure opportunities they have today.

4) the testees answer the question and get the score (three options of answers are presented):

Nowadays, the problem of public health has become (become) incredibly important worldwide. Many people all over the world forget about their physical development and health because of a whole new set of leisure opportunities they have today.
In the process of editing, the testing instructor may also apply one of the additional codes to signify the level of complexity of each question (up to 3), so that the test can be administered as an adaptive quiz: it starts with the easiest question, and if the answer is correct, the next question is more difficult, but if the answer is incorrect, the next question stays at the lowest level of complexity. This is repeated at each step: the success in answer leads to the increase in the complexity of the next question, and the failure, to the lower level of complexity. Correspondingly, the correct answers have different weights depending on the level of the question.

The following data can show the efficiency of RETM-produced tests administered to two groups of first-year students of the Higher School of Economics:

- Number of tests - 2
- Number of questions in each test: 1 - 50, 2 - 60;
- Number of students - 54
- Testing areas - tense choice, tense form, voice choice, infinitive construction, gerund construction, participial construction
- Range of correct answers: 46% to 81% with the average success of 66%
- Increase in the success rate in diagnostic samples from pre-testing to post-testing in 67% of students, the same rate of success in 13% of students, and decrease in 20%.
- Number of questions deleted at the editing stage - 12 of 122.
- Time for editing 122 questions - about 50 minutes.

With the help of such semi-automatically generated tests English instructors can make sure that their students should not repeat the mistakes that they themselves have previously made in their written works. Moreover, the tests can be administered as a specially designed course of tests from simpler to gradually more sophisticated. Finally,
the courses of tests can be a part of work in class or of independent learning trajectory, while the instructor gets the necessary statistics in any mode of administration.

The tests produced by RETM allow English instructors to save much time required for the process of collecting and organizing testing materials and grading the results of the tests, even though instructors do have to carry out some preliminary evaluation in the semi-automatic regime at the stage of finalising the testing questions. In addition to being a useful tool for the teaching-learning process, RETM tests also provide the area for more research into language acquisition and make the teaching process much more custom-made.

References


This paper examines the extent to which *you know* is used as a discourse marker by Saudi native speakers of Arabic. The Saudi learner corpus (henceforth LINDSEI –AR) employed in this study is an official component of the Louvain International Database of Spoken English Interlanguage (LINDSEI) and as such, the design is determined by the general LINDSEI model. This corpus is the first part of LINDSEI representing English used by Saudi learners. It consists of 50 informally recorded interviews in English between Saudi students studying English as foreign language. In total the corpus comprises almost 68,100 words of interviewee speech. The participants in this study are all third and fourth year university students (males and females, aged 20–25 years) majoring in English linguistics/translation. The level of study is a criterion for this corpus (De Cock, 2004; Gilquin, 2008; Gilquin et al, 2010).

Not only is the use of *you know* in this learner corpus examined from a quantitative and a qualitative perspective, it is also compared with native speaker corpus, the Louvain Corpus of Native English Conversation (LOCNEC) (De Cock 2004). In fact, following the standard format of LINDSEI gives this learner corpus the advantage of using native speaker of English data, LOCNEC, compiled by De Cock (2004), which also used the same structure as LINDSEI. Native speakers in LOCNEC were British students majoring in English language and/or linguistics at the University of Lancaster.

Qualitatively, the functions of *you know* are analysed with a moderately corpus-based approach, implementing functions of *you know* from previous studies (e.g. Brinton, 1996; Fraser, 1990; Fraser, 1999; Fraser, 2006; Schiffrin, 1987). In order to distinguish *you know*’s discourse marker functions from its other uses, key characteristics such as optionality, orality, multi-functionality identified in previous researches are considered (e.g. Brinton, 1990; Brinton, 1996; Schiffrin, 1987; Schourup 1999). Regarding the quantitative part of the study and the frequency of discourse markers, the tokens are counted manually and then double-checked via a computer search. The results of frequency of *you know* is normalized, that is, standardized according to a consistent text length of 10,000 words. With this information, the number of *you know* tokens is calculated and multiplied by 10,000 and divided by the total word count.

The investigation shows that the six discourse marker functions of *you know*, following Brinton’s (1996) classification three textual and three interactional, are used by Saudi speakers and native speakers as can be seen in Table 1.1 and Figure 1.1. Most of the functions are used by Saudi speakers more than native speakers except *you know* for lexical or content search. The most frequent function in both corpora is *you know* for speaker/hearer shared knowledge. The difference between the two corpora for this function achieved statistical significance p <0.0001.
Table 1.1 Distribution of you know in the LINDSEI-AR and LOCNEC (frequencies per 10,000 words in parentheses)

<table>
<thead>
<tr>
<th></th>
<th>Search</th>
<th>Restart &amp; Repair</th>
<th>Explain</th>
<th>General Knowledge</th>
<th>Shared Knowledge</th>
<th>Narrative</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLC (68,100 words)</td>
<td>52 (8)</td>
<td>49 (7)</td>
<td>11 (2)</td>
<td>33 (5)</td>
<td>161 (24)</td>
<td>52 (8)</td>
</tr>
<tr>
<td>LOCNEC (68,000 words)</td>
<td>63 (9)</td>
<td>32 (5)</td>
<td>20 (3)</td>
<td>20 (3)</td>
<td>91 (13)</td>
<td>47 (7)</td>
</tr>
</tbody>
</table>

Figure 1.1 Distribution of DM functions of you know in LINDSEI-AR according to gender

Regarding gender in the learner corpus, similar to results in Macaulay (2002), it is found as in Table 1.2 and Figure 1.2 that the female speakers used a higher rate 'raw frequency' of you know with 200 tokens or 65 per 10,000 words compared to the male speakers who used it 228 times, or 61 per 10,000 words. The frequency of the non-discourse marker and the unclear instances was higher for the male speakers than the females. However, neither frequency difference was statistically significant.

Table 1.2 Number of tokens of you know in the LINDSEI-AR according to gender (frequencies per 10,000 words in parentheses)

<table>
<thead>
<tr>
<th></th>
<th>Total Frequency</th>
<th>Non-DM use</th>
<th>DM use</th>
<th>Unclear</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male (37,108 words)</td>
<td>228 (61)</td>
<td>23 (6)</td>
<td>183 (49)</td>
<td>21 (6)</td>
</tr>
<tr>
<td>Female (30,900 words)</td>
<td>200 (65)</td>
<td>14 (5)</td>
<td>175 (57)</td>
<td>11 (4)</td>
</tr>
</tbody>
</table>
The analysis of the discourse marker use of *you know* shows that the gender has some effect. Even though the number of tokens is quite small, it is worth discussing. Table 1.3 and Figure 1.3 illustrate in detail the distributional pattern between male and female speakers. For both gender groups, *you know* is used most frequently as a speaker/hearer shared knowledge marker. The female speakers used 95 tokens, or 31 per 10,000 words in their speech, whereas the males used 66 tokens, or 18 per 10,000 words. This difference is highly significant (p < 0.0001). The female speakers used *you know* for speaker/hearer shared knowledge more than the total of all other functions (80 versus 95). My results are similar to those of Macaulay (2002) who found that women are more likely to use *you know* than men. In another study by Stubbe and Holmes (1995), *you know* was found to be more frequent in the speech of the young male than in the speech of the young female, both from a working class background. However, because LINDSEI does not provide information about the social background of the participants, the results of this study are not compared to those of Stubbe and Holmes (1995).

**Table 1.3 Functions of you know in the LINDSEI-AR according to gender**

<table>
<thead>
<tr>
<th></th>
<th>Search</th>
<th>Restart &amp; Repair</th>
<th>Explain</th>
<th>General Knowledge</th>
<th>Shared Knowledge</th>
<th>Narrative</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Male</strong></td>
<td>41 (11)</td>
<td>15 (4)</td>
<td>6 (2)</td>
<td>17 (5)</td>
<td>66 (18)</td>
<td>38 (10)</td>
</tr>
<tr>
<td>(37,108 words)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Female</strong></td>
<td>11 (6)</td>
<td>34 (11)</td>
<td>5 (2)</td>
<td>16 (5)</td>
<td>95 (31)</td>
<td>14 (5)</td>
</tr>
<tr>
<td>(30,900 words)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
In fact, there are considerable individual differences in the LINDSEI-AR in terms of the rate of occurrence of the discourse marker you know. There are some speakers who used you know at a markedly higher rate (one speaker 30 tokens and another speaker 34 tokens) in an average of 1,400 words per recording, which is a clear contrast to the other speakers’ use which ranged from two to maximum 20 tokens. On the other hand, the analysis of the native speakers data showed that the highest frequency of you know use was 35 tokens in one interview, interviewee E11. However, the average total number of words in E11 interview (B turns) was 2,800 words. So, in this situation, I argue that high use of you know which occurred in the LINDSEI-AR is not necessarily due to the lack of fluency of speakers but rather occurred as ‘speech habit’ (Östman, 1981, p. 27) or what Macaulay calls ‘part of the speaker’s discourse style’ (2002, p. 765). Macaulay (2002) points out that the discourse marker you know does not appear to mark ‘assumptions of shared knowledge, but rather to form part of the speaker’s discourse style and the rhythmic organization of utterances’ (2002, p. 765). Even if Macaulay’s argument is valid, it is still necessary to classify tokens of you know which conform to shared knowledge as a discourse marker use and this is what has been done in this study.

References


Average surprisal of parts-of-speech
Hannah Kermes and Elke Teich (Universität des Saarlandes, Germany)

We present an approach to investigate the differences between lexical words and function words and the respective parts-of-speech from an information-theoretical point of view (cf. Shannon, 1949). We use average surprisal (AvS) to measure the amount of information transmitted by a linguistic unit. We expect to find function words to be more predictable (having a lower AvS) and lexical words to be less predictable (having a higher AvS). We also assume that function words’ AvS is fairly constant over time and registers, while AvS of lexical words is more variable depending on time and register.

Our assumptions are based on well known differences between lexical words and function words with respect to frequency, word length, number (open vs. closed class) and information content, lexical words being the main carriers of meaning (Biber et al., 1999). Besides, Piantadosi, Tily, and Gibson (2011) show that average information content is a better predictor for word length than frequency. According to Quirk et al. (1985, 72) the choice is larger in typical contexts of lexical words than of function words and Linzen and Jaeger (2015) provide evidence that the number of choices in a particular context affects the predictions of people for upcoming syntactic construction.

As an example we look at the development of scientific English. We assume that due to specialization, scientific texts exhibit greater encoding density over time Halliday (1988); Halliday and Martin (2005), i.e. more compact, shorter linguistic forms are increasingly used, in order to maximize efficiency in communication. One feature of linguistics densification is the extensive use of lexical words (often approximated by lexical density). Thus, we expect to see differences in the AvS of lexical words in scientific writing over time and with respect to general language.

Data and Methodology
To test our assumption about the constancy/variability in AvS of lexical words and function words over time and registers, we focus on the period of Late Modern English using two data sets the Royal Society Corpus (RSC, Kermes et al., 2016) and the Corpus of Late Modern English Texts, version 3.0 (CLMET, Diller et al., 2011).

The RSC is a historical corpus of written scientific English based on the first two centuries of the Philosophical Transactions of the Royal
Society of London (1665–1869) and comprises approx. 35 million tokens. With its long and continuous history the journal provides a very good basis for diachronic analysis of English scientific writing. The RSC is annotated for lemma and parts-of-speech using TreeTagger (Schmid, 1994, 1995). CLMET is a register-mix corpus with a similar size, time span (1710–1920) and comparable annotation (Penn Treebank tag set, Marcus et al., 1993).

As a measure of surprisal, we use a model of AvS, i.e. the average amount of information a word encodes in number of bits, calculated as

\[
AvS(\text{unit}) = \frac{1}{|\text{unit}|} \sum_{i} - \log_2 p(\text{unit}|\text{context}_i)
\]

i.e. the (negative log) probability of a given unit (e.g. a word) in context (e.g. its preceding words) for all its occurrences (cf. Genzel & Charniak, 2002). In general, surprisal Levy (2008) captures the intuition that the less probable a linguistic unit is in a given context, the more surprising or informative that unit will be and the more bits are needed to encode it (and vice versa). This allows us to investigate the differences between lexical words and function words with respect to information content and predictability synchronically and diachronically looking at the distribution of AvS for each part-of-speech group including the range/spread of AvS, (relation of) mean and median. The AvS values for each token are annotated in the corpora for an easy access.

We extract all words, excluding non-word items from each corpus with information about its parts-of-speech (UPenn tagset), AvS and time period. For a better abstraction we group the parts-of-speech into function words (article, preposition, pronoun, modal, conjunction and the auxiliaries be and have), lexical words (noun, adjective, verb, adverb), and other.

### AvS of parts-of-speech

Figure 1 displays the distribution of AvS values for parts-of-speech in the RSC. Function words are to the left of the diagram (article, preposition, pronoun, modal, conjunction and the auxiliaries be and have), lexical words to the right (noun, adjective, verb, adverb).

Generally, we can observe the following differences in the distribution of AvS for lexical words and function words. Lexical words have an almost equal mean and median, the distribution has a large spread/range and is mostly symmetric with a relatively flat curve. Function words have a lower mean than lexical words, the median is often lower than the mean with distinct peaks mostly to the lower end.
Function words behave more diverse than lexical words. *Articles, prepositions* and *conjunctions* are positively skewed with the major peak shifted to the lower end of the distribution and the median being lower than the mean. The pronounced peak of *prepositions* to the far lower end is related to the preposition *of* in complex noun phrases. *Pronouns, modals* and the *auxiliaries* show characteristics of both function words and lexical words. *Pronouns* have a mostly symmetric distribution, mean and median being almost equal with a pronounced peak around the median. *Modals* and *auxiliaries* have flatter curves than the other function words, with less distinct peaks. The mean of *modals* is high in comparison to the other function words.

We compare these distributions to the AvS of parts-of-speech in CLMET (Figure 2) to see whether our findings are specific for scientific language or whether they have a more general character.

In general, we can observe a similar picture. There are differences between function words and lexical words with both groups exhibiting more or less the same general properties. A closer look reveals differences between CLMET and the RSC for specific parts-of-speech:

- the distribution of *nouns* is less symmetric in CLMET with a distinct peak to the higher end.
- the curves of *modals* and *auxiliaries* is more pointed with at least two peaks.
- the positive skew of *articles* and *prepositions* is less pronounced

The tendencies can be related to linguistic complex structures (such as complex noun phrases) being more common and thus more predictable in the RSC than in CLMET.
If we now look at the diachronic development of AvS of parts-of-speech, we can observe that the AvS remains relatively stable over time in CL-MET, mean, average and shape of the distributions hardly change (cf. Figure 3). In the RSC, however, we can observe small changes for some of the parts-of-speech. For typical modifiers such as adverbs, modals as well as for pronouns AvS increases slightly. For articles, prepositions and nouns as well as for verbs and the auxiliaries be and have AvS decreases.

Overall, the range of the distribution increases for all parts-of-speech and there is a general trend to develop peaks to the lower end of the
The differences that we observe in synchronic comparisons of CLMET and the RSC get stronger over time. In other words, scientific writing gets more distinct from “general language” over time with respect to the AvS of lexical and function words.

References


The conceptualization of music in semantic frames based on word sketches
Ana Ostroški Anić (Institute of Croatian Language and Linguistics, Croatia) and Sanja Kiš Žuvela (Music Academy, University of Zagreb, Croatia)

It has been standard practice in terminology management to define the categories of specialized knowledge as complex conceptual networks in which concepts are connected by ontological relations. With the emergence of Frame-based Terminology (Faber, 2012) that applies Frame Semantics (Fillmore, 1985; Fillmore and Atkins, 1992), it has become evident that specialized knowledge categories can be more adequately defined by using semantic frames as conceptual structures that allow for a precise description of frame elements and their corresponding lexical units.

Apart from few suggestions of automatic utilization of semantic relations (Materna, 2014; McCarthy et al., 2015; León-Araúz, San Martín, and Faber, 2016), the construction of semantic frames still relies heavily on semiautomatic corpus based methods and manually analysed data. A dynamic terminological description of a specialized domain that takes into account conceptual metaphors, variation and polysemy as integral components of specialized knowledge needs to be based on a rigorous linguistic analysis, but can benefit greatly from an automatic linguistic description such is offered by the word sketches (Kilgariff et al., 2014).

The paper gives an analysis of several semantic frames in the domain of music (e.g. PITCH, TONALITY, TEMPO) that are constructed based on the word sketches extracted from an English corpus of scientific papers and books in music theory. The analysis is carried out in order to test the reliability of the linguistic information in word sketches as opposed to the information extracted by a common terminological practice of analysing concordances and knowledge-rich contexts. Special attention is given to figurative terminological units and metaphorical mappings in the frame elements.

Semantic frames are first constructed based on the extracted word sketches of most frequent terms. Frame elements are defined following the FrameNet methodology, which has been adapted, where necessary, to better reflect the nature of specialized knowledge categories (Faber, León-Araúz, and Prieto Velasco, 2009; Ostroški Anić, 2015). To evaluate the relevance of linguistic information in the frames, key terms used for the extraction of word sketches are analysed in corpus concordances, and additional relevant information is marked accordingly.

The research is conducted within the project Problems of basic contemporary musical terminology in Croatia (CONMUSTERM), and follows on a model of analysis of figurative specialized language based on application of a semiautomatic method of metaphor detection, and the use of semantic frames for a description of metaphorical mappings between the categories of general and specialized knowledge.
References


Animals and their places in news discourse: insights from cross-linguistic and diachronic perspectives
Anda Drasovean (King’s College London, UK) and Emma McClaughlin (Lancaster University, UK)

The discursive representation of animals in contemporary British English has been investigated by a recent project across a range of spoken and written discourse types. Two adjoining studies attached to the project focused on the representation of animals in the news. The two projects differed in that one took a cross-linguistic approach whilst the other examined discourse about animals from a diachronic perspective. The data for the diachronic project comprised three sub-corpora of letters and articles about red and grey squirrels, badgers, and hedgehogs published in The Times newspaper between 1785 and 2005. The cross-linguistic project compared how five case-study animals (bears, pigs, dogs, horses, and cats) are represented in a purpose-built corpus of contemporary (2012-2014) online news stories from Romania and the UK.

Both studies employ a corpus-assisted discourse analytical approach using a range of procedures such as keyword, collocation, word clusters, word sketch, and concordance analyses. Close analysis of representative texts from the corpora was also carried out to add qualitative depth. A number of common themes and patterns were identified between the two studies including animal population control, disease, and conservation.

One common salient pattern identified by both studies is that of spatiality. Disciplines such as animal geographies have highlighted the key role that spatiality plays in the representation of animals. For instance, the places inhabited by animals, whether real or imaginary, are fundamental to how animals are classified and conceptualised (Philo & Wilbert, 2000). A dominant paradigm in defining human-animal relationships is the distinction between human spaces and animal spaces, more commonly referred to as the nature-culture boundary. The language used to discuss animals in the analysed corpora illustrates different ways in which the nature-culture boundary is reinforced or renegotiated in news stories.

In the diachronic study, wildlife species that cross the nature-culture boundaries and enter what are considered to be “human” spaces are generally framed as problematic; however the corpus also revealed limited examples of how animals entering such spaces are not problematic if their presence is beneficial in some way. For example, hedgehogs eating garden pests are described as “invaluable friends” (1954); “the gardener’s friend” (2005); and “garden favourites” (2005). The analysis of the Romanian corpus suggests that, in the Romanian context, wildlife species such as bears are most frequently described as problematic. For example, news stories about bears often report cases in which they cross into human territory and attack humans or domestic animals. In these stories, the frequently recurring phrase “marginea padurii” (the edge of the forest) defines and reinforces the boundary between human and animal territory.

Using this, and similar illustrative examples, we discuss the benefit of a combined cross-linguistic and diachronic approach to the corpus-assisted discourse analysis of language about animals. We propose that the cross-linguistic and
diachronic lenses can be used to complement the investigation of any social, cultural, and political issue for which a synchronic, monolingual corpus-linguistic approach might traditionally be used.

References

Over the past decade, research into English as a lingua franca (ELF) has steadily developed into a thriving field. ELF has been studied from the perspective of pronunciation, but also lexis, grammar and pragmatics (Dewey, 2007; Jenkins, 2007; Ur, 2010; Walker, 2010; Kaur, 2011; Seidlhofer, 2011; Önen, 2014). Our research proceeds from two basic assumptions: first, that English does not belong to native speakers, since there are now more non-native than native speakers of English (Graddol, 2006; Seidlhofer, 2011) and second, that ELF is not a language as such, but rather “a means of communication not tied to particular countries and ethnicities, a linguistic resource that is not contained in, or constrained by, traditional (and notoriously tendentious) ideas of what constitutes ‘a language’” (Seidlhofer, 2011, p. 81). The latter idea is further echoed in Kohn (2015a; 2015b) who argues for a reconciliation between ELF and English language teaching, which still dominantly tends to focus on the correct use of standard English rather than the communicative needs of students.

Although ELF has obtained world-wide recognition, no excessive research on its lexicogrammatical features has been carried out in Estonia, with the exception of Soler-Carbonell (2014; 2015), who has explored the sociolinguistic aspects of ELF and university students’ attitudes towards ELF in academic settings. Our study is therefore innovative as it broadens the scope of research on ELF by looking at a language that is typologically different from the (predominantly Indo-European) languages usually discussed in the studies carried out within this field. By comparing the data of Estonian ELF speakers to the data of other languages that have served as input for describing the characteristics of ELF, our study makes a crucial contribution towards validating the generality of the proposed characteristics of ELF.

For this purpose, 9 semi-structured interviews were recorded with Estonian ELF speakers at B1-C2 levels (all university students). All participants reported to be native speakers of Estonian (8 women and 1 man; average age 22.4 years). The interviews were transcribed using the free software EXMARaLDA and the VOICE conventions for transcription (https://www.univie.ac.at/voice/). In addition to the qualitative analysis of the data, a mixed-effects logistic regression model is fitted to the data in order to filter out any biases likely to be due to individual differences. Our data reveals some of the features identified by Seidlhofer (2004, p. 220), Dewey (2007) and Önen (2014), such as omitting and inserting articles, inserting, omitting and substituting prepositions and using verbs with high semantic generality in innovative collocations.

As to the practical outcomes, the findings from our mini-corpus could serve as reference material to teachers of English in Estonia who wish to increase their own and their students’ awareness of ELF, since this is the foundation that should be laid before any specific features could be taught (see also Sifakis, 2014). We also hope to develop the corpus into a larger one by conducting more interviews and possibly also including other native languages, including speakers of English as a native language in order to facilitate direct comparisons between different groups of English speakers.
References


The HeidelGram project, based at the English Department of Heidelberg University, has a twofold aim. Firstly, it makes an essential contribution to historical grammar studies by compiling, making available and analysing a representative, XML-annotated, 10-million-word corpus of historical English grammar books from the 16th to the 19th centuries. Secondly, it introduces state-of-the-art network analysis into historical corpus linguistics in order to considerably extend the set of concepts and methods applied in historical linguistics and corpus linguistics, and to exemplarily implement and analyse various kinds of networks, such as a network of grammarians or of lexemes of Sprachkritik to examine complex discursive manifestations of the modern framework of verbal hygiene (Cameron 1995) in long-term diachrony.

The combination of corpus-based historical linguistics (see Jucker and Taavitsainen 2014: 4) and network analysis (see Freeman 2004, White 2011) has been rather uncharted territory. Based on network and graph theories (see, for instance, Jungnickel 2013), the analysis of social networks has become a key technique in sociology. Since its emergence in the early 20th century (see Freeman 2004), the approach has been adopted by many other disciplines, such as communication and media science, information science, biology, geography, and economics. Recently, network analyses have been conducted using social media data, literature, or drama (e.g. Elson et al. 2010, Agarwal et al. 2012, Moretti 2013), but so far there have not been any network-based linguistic studies on historical non-fictional texts.

This pilot project constitutes the first part of a series of network analyses of full-text historical English grammar books. Since previous research on grammar writing (e.g. Wischer 2003, Tieken-Boon van Oostade 2008, Anderwald 2016) was mostly restricted to shorter periods of time, specific linguistic phenomena or single authors or grammars, and was not corpus-based, this is the first study that begins to analyse grammar books from a large-scale perspective.

The present study investigates the varied connections between 19th-century grammarians by analysing references authors make to other 19th-century grammars and grammarians. Applying Mehler's concept of citation networks (Mehler 2008: 356ff) and White's notion of scholarly networks, references are understood as "record[s] of who has cited whom within a fixed set of authors" (White 2011: 275). The network will show the in-citations, i.e. referenced grammarians, and the out-citations, i.e. the grammar books that contain these references (ibid.). The systematic examination of grammarians' references in 19th-century grammar writing serves as a starting point to identify influential and isolated grammarians, and discursive strategies which show how grammarians deal with alleged authorities in the field.

A pilot corpus of 19th-century British grammar books (40 texts, ca. 2.6 mio. words) forms the basis for this pilot network analysis. It contains the most well-known and widely distributed grammars of the 19th century (cf. Leitner 1986, 1991, Linn 2006, Michael 1987, Görlach 1998), such as Crombie (1802), Cobbett (1818), Sweet (1892/1898) and Nesfield (1898), as full texts in digitised form. Main criteria
for text selection are numbers of published editions, their distribution, and common use, as found in book catalogues and secondary literature on grammar writing.

A first manual search in the grammar books, supplemented by those names that are usually considered the most famous and influential authors of their time (cf. Dons 2004, Finegan 1998, Görlach 1998, Linn 2006, Schmitter 1996, Tieken-Boon van Ostade 2008, Wolf 2011), results in a list of English and foreign grammarians that are referred to. This list of about ninety referenced authors, applied to the pilot corpus, yields all references made to other grammarians.

After manually deleting false positives, such as references to other people with the same name, the concordance list is turned into a multi-dimensional matrix and into different kinds of networks for chronological visualisation.

The dyadic ties between authors are examined quantitatively, i.e. in terms of the number of references, and qualitatively, i.e. by classifying different kinds of references, e.g. quotation, approval of approaches to grammar, the citing of authorities, and various forms of criticism. Approval, for instance, is "I concur with Baker in considering ..." (Crombie (1802) on Baker (1724)), whereas examples of criticism are "Mr. Cobbett has mistaken the real causes of defective arrangement" (Doherty (1841) on Cobbett (1818)) and "Mr. Harris could declare nothing more incomprehensible to man's understanding, than [...]" (Murray (1847) on Harris (1751)). Authors might also adopt other grammarians' views, as in Murray (1847) referring to Crombie (1802), who criticises Lowth (1762): "If, as Doctor Crombie informs us, [...] Doctor Lowth violated the simplicity of our language, [...], that is a sufficient reason for not receiving his grammatical opinions, without first strictly examining their truth, and next their applicability to the English language". By classifying the various kinds of references, we are able to show different and changing attitudes towards other grammarians' approaches to grammar, and to discuss substantial implications for the development of the genre and 19th-century language practices.

The network of references further reveals paradigm shifts in grammar writing, systematically charting particularly the rise of descriptive grammars after the predominance of prescriptivism and, at the same time, critically reflecting on fixed assumptions of what is known to be 19th-century 'prescriptive' and 'descriptive' grammar writing. Preliminary results show, for example, that the rejection of prescriptivism is by no means a uniform process and that prescriptive aspects in grammar books never vanish completely. Moreover, grammarians do not use references uniformly. While some authors usually quote other grammarians without stating their own opinions, others show criticism and disapproval openly. Interestingly, however, those self-opinionated authors who criticise other grammarians most frequently are themselves never referred to in other grammar books.

References


The objective of this study is to investigate linguistically the ways in which systems of meaning are constructed in British newspaper editorials to represent 'self' and 'other' through currently circulating discourse of 'war on terror' in ideology formation. The study explores the newspaper editorials' role in shaping stereotypical images and ideologies as a result of the events of September 11, 2001, which are still on their way in the form of 'war on terror' discourse. In this research, a corpus-based approach to critical discourse analysis, rather more specifically, corpus-based approach to ideological analysis is used for data analysis. The study employs Van Dijk's concept of ideological square (1993, 1995, 1998 and 2008) as its theoretical framework. The Ideological square, comprises the semantic macro strategies developed by Van Dijk which provides the features for positive self-presentation and negative other-presentation. This ideological square emphasises the positive 'Us' and de emphasises the positive 'Them'; similarly, it emphasises the negative 'Them' and de emphasises the negatives 'Us'.

The corpus of the research consists of retrieving all the editorials/opinion-editorials/leading articles about 'War on Terror', from two British newspapers 'The Guardian' (TG-corpus) and 'The Times' (TT-corpus) from September 11, 2001 to December 31, 2011. The editorials, chosen for analysis, have been carefully collected and sorted out to make sure that only editorials/leading articles are included from 'Lexis/Nexis' and 'ProQuest' databases. Wordsmith Tools (2015), Sketch Engine (2010, 2014) and WMatrix softwares (2009)/Web portals have been used for Corpus annotation. Both the corpora were tagged semantically and grammatically using USAS and CLAWS from Lancaster's platform.

Background

After the September 11 attacks in New York and Washington the United Kingdom expressed its sympathy towards the United States and pledged to assist in any way possible in the new 'war against terrorism'. The newspapers across the United Kingdom framed the main leads in significant ways.. The main leads of the elite national newspapers ran: War on America\(^1\), A declaration of war\(^2\), Day that changed the world\(^3\), Is this the end of the world?\(^4\), War on the world\(^5\), Apocalypse\(^6\), 10.02 am September 11 2001\(^7\), Declaration of War\(^8\).

Surprisingly, the British newspapers' rhetoric in the wake of September 11 attacks, constituted the chaotic imagery of war: war, terror, declaration, attack, End times and Apocalypse. Here, the imagery focused on the 'war' with images of collapse, devastation, anguish, outrage and nightmare; images of death, doomsday and apocalypse.

The frequency of editorials that present 'war on terror' discourse, published overtime from 9/11 to 31\(^{st}\) of December, 2011 in The Guardian and The Times are shown in the following Table:

1 The Daily Telegraph
2 The Guardian
3 The Sun
4 Daily Star
5 The Mirror
6 Daily Mail
7 The Times
8 Daily Express
The TG-corpus comprises of all the editorials that were published from 9/11, 2001 to December 31, 2011 in The Guardian on the theme of ‘war on terror’. The overall frequency of TG-corpus is 486,279 words. The TT-Corpus comprises of all the editorials published from 9/11, 2001 to December 31, 2011 in The Times on the topic of ‘war on terror’. The overall frequency of TT-corpus is 659,711 words.

**Ideological Square: Positive Self Presentation**

To find out how ‘self’ is represented in British newspapers the first right hand collocates/ semantic preferences are derived from the TG-corpus. In the following table, the most of the semantic preferences of the British revolve around the vocabulary of politics, law and order, foreign policy, war and community; presenting the ‘British’ in overall a positive manner.

<table>
<thead>
<tr>
<th>Newspaper</th>
<th>Year</th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Guardian</td>
<td></td>
<td>84</td>
<td>98</td>
<td>85</td>
<td>74</td>
<td>65</td>
<td>115</td>
<td>74</td>
<td>78</td>
<td>59</td>
<td>49</td>
<td>59</td>
<td>840</td>
</tr>
<tr>
<td>The Times</td>
<td></td>
<td>127</td>
<td>141</td>
<td>107</td>
<td>126</td>
<td>130</td>
<td>112</td>
<td>85</td>
<td>74</td>
<td>52</td>
<td>49</td>
<td>81</td>
<td>1084</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>211</td>
<td>239</td>
<td>192</td>
<td>200</td>
<td>195</td>
<td>227</td>
<td>159</td>
<td>152</td>
<td>111</td>
<td>98</td>
<td>140</td>
<td>1,924</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Semantic Preference/collocates of ‘British’ in TG-corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>troops, government, Muslims, people, forces, prime, citizens, public, nuclear, state, soldiers, policy, foreign, governments, military, army, intelligence, politics, ministers, ambassador, support, bill, officials, passports, lawyers, law, society, politicians, commanders, authorities, nationals, Muslim</td>
</tr>
</tbody>
</table>

The rhetoric of Western ideals of Justice, liberty, freedom and democracy is also frequently present in the British broadsheets. This rhetoric inspires these ideals to present the ‘self’ in a civilised way and the ‘other’ in highly negative way. The overall policy of TG is of a critique of the American war mongering and glorification of its ideals of freedom, liberty, democracy and justice. In the table below, the most of the semantic preferences of ‘British’ in TT-corpus are very much similar to TG-corpus presenting the ‘British’ in overall a positive manner.

<table>
<thead>
<tr>
<th>Semantic Preference/collocates of ‘British’ in TT-corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>people, airways, soil, special, citizen, officials, national, justice, economy, embassy, interests, American, Muslim, police, security, men, courts, commanders, soldier, airports, mainstream, support, transport, armed, army, history, voters, policy, commercial, company, mosques, policing, politics, Islam</td>
</tr>
</tbody>
</table>

TT-corpus frequently refers to American ideals of civil liberty, justice, freedom and democracy as the Western values throughout the discourse for its appropriation and rationalisation.

**Ideological Square: Negative Other Presentation**

The empirical findings show that in both The Times and The Guardian corpora the negative ‘other’ presentation is unanimously and consistently revolves around the terrorists, militants, extremists, militias, fundamentalists and rebels; and interestingly all these words are the strong collocates of Islam, Islamic and Islamists in the corpus.

The attitudes of British press overall “towards Muslims in the United Kingdom have not been positive” (Baker, 2013, p. 1). Even the biased trend of the British media can be witnessed in Paul Baker’s study of the representation of Islam and Muslims in the British newspapers, where he notices that “It was
particularly difficult to make distinctions between the final four categories (Islamic political groups, terror/extremism, crime and conflict) as these concepts overlapped” (2013. p. 74).

Table below shows the SemTag collocates of Islamist in TG-corpus. The top collocate in the corpus (E3-) relates the Islamists to violence, extremism and show them in the state of being angry people (S2). G1.2 places them in the category of ‘politics’ signifying their political aims. The TG discourse voices a concern of dislike (E2-) and portrays them as a hindrance (S8) to the Western civilisation. (S5+) shows them as a member of ‘group’, classifying them in the ideological mode of ‘fragmentation’.

Table: **SemTag collocates of Islamist in TG-corpus**

<table>
<thead>
<tr>
<th>Collocation</th>
<th>MI</th>
<th>T-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Islamist E3-/G1.2/S2</td>
<td>7.82</td>
<td>2.23</td>
</tr>
<tr>
<td>Islamist S8-/E2-</td>
<td>5.3</td>
<td>1.69</td>
</tr>
<tr>
<td>Islamist G3/S5+</td>
<td>4.89</td>
<td>1.67</td>
</tr>
<tr>
<td>Islamist E3-/Q2.2</td>
<td>4.83</td>
<td>1.93</td>
</tr>
<tr>
<td>Islamist G2.1-</td>
<td>4.73</td>
<td>3.19</td>
</tr>
<tr>
<td>Islamist S1.1.3+</td>
<td>4.67</td>
<td>1.92</td>
</tr>
<tr>
<td>Islamist S9/S2</td>
<td>4.66</td>
<td>1.66</td>
</tr>
<tr>
<td>Islamist S2.1-</td>
<td>4.51</td>
<td>2.34</td>
</tr>
<tr>
<td>Islamist S5+</td>
<td>4.06</td>
<td>2.66</td>
</tr>
<tr>
<td>Islamist S5+c</td>
<td>3.91</td>
<td>2.09</td>
</tr>
<tr>
<td>Islamist C1-</td>
<td>3.62</td>
<td>1.84</td>
</tr>
<tr>
<td>Islamist Z2</td>
<td>1.4</td>
<td>2.06</td>
</tr>
</tbody>
</table>

In TG-corpus, word Islamist is used 82 times; the table below shows the nouns modified by ‘Islamist’:

Table: **Nouns modified by adjective ‘Islamist’ in TG-corpus**

|terrorism, terrorists, militants, extremism, movement, parties, groups, terror, militias, opposition, fundamentalists, group, rebels, organisation|

The table below shows the SemTag collocates of Islamist in TT-corpus. Here, the results of the collocational analysis are almost similar to those of TG-corpus, (E3-/G1.2) being the top collocates.

Table: **SemTag collocates of ‘Islamist’ in TT-corpus**

<table>
<thead>
<tr>
<th>Collocation</th>
<th>MI</th>
<th>T-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Islamist E3-/G1.2/S2</td>
<td>7.11</td>
<td>4.96</td>
</tr>
<tr>
<td>Islamist E3-/G1.2</td>
<td>7.05</td>
<td>3.29</td>
</tr>
<tr>
<td>Islamist X5.2+/+++/S2</td>
<td>6.64</td>
<td>2.8</td>
</tr>
<tr>
<td>Islamist G1.2/S2</td>
<td>5.73</td>
<td>7.54</td>
</tr>
<tr>
<td>Islamists A2.1-</td>
<td>5.58</td>
<td>1.7</td>
</tr>
<tr>
<td>Islamist G2.1-</td>
<td>4.42</td>
<td>6.4</td>
</tr>
<tr>
<td>Islamist G2.1-/S2</td>
<td>4.38</td>
<td>5.47</td>
</tr>
<tr>
<td>Islamist S1.2.1-</td>
<td>4.17</td>
<td>1.89</td>
</tr>
<tr>
<td>Islamist O4.6+</td>
<td>4.16</td>
<td>1.89</td>
</tr>
<tr>
<td>Islamist G1.2</td>
<td>4.11</td>
<td>6.86</td>
</tr>
<tr>
<td>Islamist S8-/E2-</td>
<td>4.09</td>
<td>2.1</td>
</tr>
<tr>
<td>Islamist Q2.2/E2-</td>
<td>3.78</td>
<td>2.27</td>
</tr>
<tr>
<td>Islamist E3-</td>
<td>1.73</td>
<td>2.31</td>
</tr>
</tbody>
</table>
The \((X5.2+++/S2)\) represents the 'Islamists' in the highest (superlative degree) terms as negatively excited, energetic and interested people. SemTag \((A2.1-)\) shows them as orthodox unchanging folks. They are presented as un-friendly to the British way of life as the tag \((S1.2.1-)\) shows. It is interesting to note that \((O4.6+)\) SemTag portrays the Islamists as 'on fire' with a hot temperature metaphors. \((G1.2/X7+)\) shows that they are sketched as 'wanted' political criminals. The Islamists are presented among the lexicon of 'fear' and 'shock' as \((E5-)\) empirically shows. 'Islamist' is used 329 times in TT-corpus and table below shows the nouns modified by adjective 'Islamist' in TT-corpus:

<table>
<thead>
<tr>
<th>Nouns modified by adjective 'Islamist' in TT-corpus</th>
</tr>
</thead>
</table>

However, the representation of 'Muslim' in TT-corpus is fundamentally constructed around their ordinary everyday religious life. Table below shows the top collocates of Islam* in TT-corpus. The most of the collocates are having negative semantic prosody:

<table>
<thead>
<tr>
<th>Collocates</th>
<th>Freq</th>
<th>T-score</th>
<th>MI</th>
<th>MI3</th>
<th>log likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>extremism</td>
<td>64</td>
<td>7.967</td>
<td>7.918</td>
<td>19.918</td>
<td>602.182</td>
</tr>
<tr>
<td>extremists</td>
<td>62</td>
<td>7.821</td>
<td>7.211</td>
<td>19.119</td>
<td>513.422</td>
</tr>
<tr>
<td>Jihad</td>
<td>36</td>
<td>5.991</td>
<td>9.454</td>
<td>19.794</td>
<td>457.789</td>
</tr>
<tr>
<td>militants</td>
<td>42</td>
<td>6.444</td>
<td>7.446</td>
<td>18.231</td>
<td>362.223</td>
</tr>
<tr>
<td>terrorism</td>
<td>61</td>
<td>7.646</td>
<td>5.575</td>
<td>17.437</td>
<td>359.568</td>
</tr>
<tr>
<td>Hamas</td>
<td>39</td>
<td>6.197</td>
<td>7.035</td>
<td>17.605</td>
<td>311.553</td>
</tr>
<tr>
<td>militant</td>
<td>19</td>
<td>4.34</td>
<td>7.817</td>
<td>16.313</td>
<td>174.697</td>
</tr>
<tr>
<td>groups</td>
<td>25</td>
<td>4.934</td>
<td>6.247</td>
<td>15.535</td>
<td>170.338</td>
</tr>
<tr>
<td>radical</td>
<td>18</td>
<td>4.221</td>
<td>7.631</td>
<td>15.971</td>
<td>160.05</td>
</tr>
<tr>
<td>terrorists</td>
<td>26</td>
<td>4.97</td>
<td>5.301</td>
<td>14.702</td>
<td>142.37</td>
</tr>
<tr>
<td>militancy</td>
<td>12</td>
<td>3.453</td>
<td>8.296</td>
<td>15.466</td>
<td>120.049</td>
</tr>
<tr>
<td>group</td>
<td>14</td>
<td>3.696</td>
<td>6.354</td>
<td>13.969</td>
<td>97.396</td>
</tr>
<tr>
<td>threat</td>
<td>18</td>
<td>4.105</td>
<td>4.942</td>
<td>13.281</td>
<td>89.49</td>
</tr>
</tbody>
</table>

TT's intertwining of Islam, Muslims, Islamic, Islamist and Islamism significantly 'Islamist' in the war on terror discourse is highly significant for number of reasons: Firstly, TT's reputation of presenting Islam in academia in highly negative manner is highlighted by many critics, like Poole (2002, 2006), Richardson (2004, 2007), and Baker (2013). Secondly, TT believes that though terrorists attacked World Trade Centre and Pentagon, the symbols of American global economic and military powers, but actually the Western lives with Western ideology was attacked.

TT's pre occupation with 'Islamist' and 'Israel' in the 'war on terror' discourse is ideologically significant. With the justification of the attack on Afghanistan, TT goes much further to claim legitimation of using "force against terrorists who operate against Israel too" (October 13, 2001) with an aim of thwarting 'Islamist terrorism'. TT's 'institutional racism' and an Islamophobe tendency is clear throughout the TT-corpus, as the empirical findings show. The rhetoric of 'Islamism' in TG-corpus with the pretext of immigrants, drugs, extremist, social exclusion, militant, and radical can be observed in the concordance below that is self-explanatory:
Concordance: Rhetoric of ‘Islamism’ in TG-corpus

TG and TT further represent Iraq and other suspect nations as ‘rogue states’ and as an ‘axis of evil’ and even ‘states of concern’ in the popular Bush fashion. Finally, the Muslim countries, especially the Middle Eastern ones, are positioned in post 9/11 British discourse in a relationship of complementarity and identity to each other and to terrorism as breeding grounds for terrorists, and in opposition to the civilised West.

References


A Corpus-based Approach to the Patterns of Motivated and Ideologically-invested Shifts in Translation
Long Keven Li (Macquarie University, Australia) and Xi Celia Li (Shanghai Jiao tong University, China)

A number of Chinese migrant writers have achieved success in writing in English in the past three decades. Globally, Wild Swans (1991/2003) by Chinese-British writer Jung Chang is the summit of success ever achieved by a Chinese migrant writer: it has sold more than thirteen million copies worldwide and has been translated into 37 languages. In Australia, however, another writer has achieved an even greater success: Mao’s Last Dancer (2003) by Chinese-Australian ballet dancer Li Cunxin is already a household name; its young readers’ version has been used extensively in the English curriculum. However, due to an explicitly accusative tone of Mao Zedong and the Communist Party of China that is palatable to intuition, neither book has been allowed for publication in Mainland China. Nevertheless, they have been translated into traditional Chinese and published in Taiwan, which is known for its freedom of press (Freedom House, 2016). Both translations are of an unusual situation where Source Text (ST) authors have written about their life experiences in Communist China for English readers, and subsequently contributed partially to the decision-making for the Chinese translation. In other words, the ST authors may have issued a license to allow certain translation shifts. Hence, the fact itself that the Chinese target text (TT) of Wild Swans is rather liberal came as little surprise. What is significant is that previous studies (Li, 2016; Li, 2017a) have found a large number of lexicogrammatical shifts in the Chinese translation of Wild Swans that are a manifestation of a clear manipulation of political ideology, most noticeably, a dramatic change of the representation of Mao in the Chinese translation.

This paper, as a continuation of previous studies, adopts Systemic Functional Linguistics (SFL) (Halliday and Matthiessen, 2014) as the overarching theoretical framework with a critical approach. A previous study on TRANSITIVITY and the instantial weight of clauses (Butt, 2008) in the ST and TT of Wild Swans have found that the Chinese TT is a more personalised account, and less interpretive of history when compared to the ST; in addition, the Chinese TT shows the back-staging of Mao and his diminished agency at the start of the Cultural Revolution, but the foregrounding of General Students and Red Guards, who committed the actual violence. The study on the interpersonal metafunction: MOOD, MODALITY and DEGREE OF INTENSITY confirms finding that the TT is more personalised; in addition, the Chinese translation is a more strident version of the ST: there are more and higher degrees of modality in Chinese, especially high obligation; similarly, there are more and higher degree of INTENSITY in the TT, especially adverbs of total degree, such as "completely", which have increased by 70% from the ST; moreover, shifts in MODALITY have revealed a more favourable evaluation of Zhou Enlai, but a harsher criticism of
Madame Mao. A similar study has been conducted on the MODALITY and INTENSITY in the translation of *Mao’s Last Dancer*, and, while the overall trends of shifts are similar with the translation of *Wild Swans*, the extent of shifts are nowhere as dramatic as those in the translation of *Wild Swans*; likewise, no clear patterns of ideological shifts have been identified in the translation of *Mao’s Last Dancer*.

This paper aims to further explore the latent patterns (Coulthard and Sinclair, 1975; Butt, 1983) of ideological shifts with a corpus-based approach utilising computational tools in Corpus Linguistics (CL) such as SysFan (Wu, 2000), SysConc (Wu, 2000) and LancBox (Brezina et al., 2015). This study is corpus-based in two senses. Firstly, the source texts and target targets of both *Wild Swans* and *Mao’s Last Dancer* are turned into machine-readable corpora, each of which is a large body of text, e.g., the ST of *Wild Swans* contains more than 220,000 words, considerably larger than Matthiessen’s (2006) recommended limit of 15,000 words for a specialised corpus. They are digitally ready to be examined for their lexicogrammatical patterns. Secondly, results from these texts will be compared with appropriate reference corpora, such as FLOB corpus of British English in early 1990s and LCMC corpus of Mandarin Chinese around 1991, with the aim to filter out translation universals (Baker, 1993, in Laviosa, 2002, p. 18) and typological issues from motivated selections by an author/translator. Areas to be investigated include word frequency, quantitative representation of major represented participants, word associations, modality and degree of intensity.

The previous studies were geared towards a more in-depth analysis of the ideational and interpersonal metafunctions within SFL in selected chapters. This paper aims to complement previous studies by adopting a more quantitative-based analysis of larger bodies of texts, with special attention paid to any shift in the representation and evaluation of individuals, groups and institutions representing the communist party of China. Firstly, concordance programmes such as SysConc and LancBox will be used to compare word frequency in the ST and TT to identify what has been back-staged or foregrounded on a broad front. Secondly, word associations with these organisations and people are to be examined as they are considered potentially ideologically invested; in addition, word associations with expressions of high modality and total intensity are also considered ideologically volatile, as they reveal the textual environments upon which the highest degrees of interpersonal judgements are projected; LancBox will be useful in explicating word associations. The next step is to address the thorny issue of translation universals and language typology, by comparing results hereby with the reference corpora, respectively FLOB and LCMC corpora. Furthermore, typical cases of translation shifts that are in accordance with the overall quantitative trends will be analysed qualitatively, to complement the quantitative methods, as only together can they ‘provide an explicative mosaic of the object under investigation’ (Wodak, 1996, p. 23). Lastly, findings from the abovementioned analysis will be linked to results from the ideational and
interpersonal metafunctions to highlight the cruces where different subsystems converge (Halliday, 1991). Subsequently, the accumulated ideological consequences will be explicated, and an attempt will then be made to bridge the semantic shifts with the differences in Contexts at the two ends of the translation process.

The significance of this paper is three-fold. Firstly, it demonstrates the efficacy of an SFL-CL combined approach to the study of ideology in translation. Secondly, it tests Hunston’s (2013) suggestion of explaining quantitative information obtained through corpus investigation with an SFL theory. Thirdly, it contributes to a contrastive study of English and Mandarin Chinese, especially in terms of modality and degree intensity; the approaches and feature lists that are built are highly applicable to similar studies.

References


“How will you make sure the material is suitable for children?”: User-informed design of Welsh corpus-based learning/teaching tools

Jennifer Needs (Swansea University, UK), Dawn Knight (Cardiff University, UK), Steve Morris (Swansea University, UK), Tess Fitzpatrick (Swansea University, UK), Enlli Thomas (Bangor University, UK) and Steven Neale (Cardiff University, UK)

The CorCenCC project (Corpws Cenedlaethol Cymraeg Cyfoes – The National Corpus of Contemporary Welsh) is breaking new ground, in that it is creating the first ever large-scale corpus of Welsh and using pioneering community-driven methods, and also in that pedagogical corpus tools are part of CorCenCC’s design from the very outset. The 10-million-word corpus will be a unique resource for Welsh language translators, lexicographers, publishers, policy-makers, language technology developers, researchers, and for those learning and teaching Welsh. This paper looks at the needs of this last group of end-users, and the challenges and opportunities arising from the task of developing a bespoke online pedagogical toolkit which works directly with the corpus data.

Welsh is one of the more privileged of the world’s lesser-used languages, in that it has government recognition and support, a rich literary tradition, dedicated television and radio stations, and is an important part of the education system in Wales. The Welsh language is taught as a second language in primary and secondary schools and in post-16 education, and it is also possible to attend Welsh-medium education from nursery right through to university-level. However, with no comprehensive corpus of Welsh, pedagogical materials have been based largely on intuition in terms of the grammar and vocabulary items to target, and authentic listening and reading passages have been few and far between.

CorCenCC’s design is user-informed, and consultation with Welsh teachers and tutors at all stages of education has shown how beneficial the corpus will be as a bank of authentic material. CorCenCC will include spoken, written and electronic language from all contexts where the language is used, e.g. in private, for socialising, for business and work, in education, in the media, and in public. It will include examples of news headlines, personal and professional letters and emails, academic writing, formal and informal speech, and even text messaging. The Welsh language varies considerably depending on formality, so CorCenCC will be a valuable resource for those teaching and learning about different genres and registers in Welsh.

CorCenCC will be freely available to the general public online via a generic user interface, but the same website will also include a second interface designed specifically for educational purposes. Both interfaces will work with the entire corpus data, but the educational interface is tailored for use by teachers and tutors from primary school to HE and adult education for use in their Welsh classes, and also at the pupils/students/learners themselves from GCSE-level up. During initial consultations with primary school teachers about CorCenCC’s potential role in their lessons, a concern arose which represents a tension between the material’s authenticity on the one hand, and considerations of appropriate classroom language
on the other: “How will you make sure the material is suitable for children?” This paper addresses the impact of this concern on the development of CorCenCC’s pedagogical interface.

**Profanity**

Perhaps the most obviously ‘unsuitable’ material for children is that which includes profane or offensive language. Because the corpus data will be uncensored, if teachers are to use CorCenCC as a resource in their classrooms, the special pedagogical interface must have an in-built tool to filter out corpus examples that contain such language. Teacher feedback thus far has pointed towards an online resource which might be used as a starting point for identification of unsuitable language, but it is by no means comprehensive and, given the bilingual situation in which the Welsh language exists, any proposed ‘filter’ must be capable of dealing with unsuitable vocabulary in English as well as Welsh. This may not be as simple as listing vocabulary items and their translations! Not all terms which are offensive in one language are necessarily offensive in translation. There is also the additional challenge of consonant ‘mutation’ in Welsh, whereby a word could potentially be spelt four different ways, depending on context – e.g. *clinig* (clinic), *dy glinig* (your clinic), *fy nghlinig* (my clinic), *ei clini* (her clinic).

**Subject matter**

*Clinig* is given as an example here, as it is evidently not an offensive or unsuitable word in itself. However, since CorCenCC will include spoken, written and e-language from a wide range of genres/subjects, it is likely that there will be coverage of certain topics which might be deemed unsuitable for children – discussion of personal medical conditions, for example, or perhaps discussion of alcohol consumption or sex. In addition to a ‘profanity filter’, CorCenCC will need a strategy for identifying and marking up ‘adult’ subject matter, so that corpus search results obtained via the pedagogical interface can be tailored to exclude passages from inappropriate texts. For example, with transcribed spoken language, ‘adult’ content can be identified during transcription.

**Advanced vocabulary**

Suitability for children does not only concern adult content, but also the content’s level of difficulty. Examples of business and legal language, for example, could potentially include vocabulary and concepts well beyond the expected level of a primary school pupil. Upon completion of the corpus, the CorCenCC project will be in a position to produce the very first frequency lists for Welsh. Plans are underway to implement a search result filter based on frequency, so that corpus-users can match search results to level of ability in Welsh.

**‘Incorrect’ language**

Another concern raised by teachers regarding ‘suitability’ for school children was whether the whole of the corpus content would model ‘correct’ Welsh. School
teaching in Wales is driven by the requirements of the National Curriculum for Wales, and for the Welsh language, even at primary school level, these entail prescriptive expectations in terms of language use.\(^1\) However, it is part of CorCenCC’s philosophy to describe the language that exists, rather than to present language data according to prescriptive norms. Corpus data will not be labelled as ‘correct’ or ‘incorrect’, and this could potentially limit pedagogical end-users’ engagement with the corpus. To ensure that the corpus does fulfil its potential as a learning/teaching resource, the project’s solution is to provide guidelines suggesting how Welsh teachers might best filter corpus search results to include only that content which is most likely to reflect the type of language they are trying to model.

Because of CorCenCC’s commitment to meeting the needs of end-users, and because those end-users include children, CorCenCC will face challenges not faced by many other corpus projects. However, these challenges also provide opportunities to find user-informed solutions, and to lead the way for future corpora. This paper will look at the pros and cons of some of the possible strategies for tackling ‘unsuitable’ language (of various kinds) in the corpus, and will illustrate the likely outcomes of the different approaches using Welsh language data. Finally, the paper will propose a ‘suitable language policy’, which could be adopted not only by CorCenCC but also by future corpora looking to meet the needs of younger corpus-users.

References


---

\(^1\) For example, for those in Welsh-medium education, Year 6 pupils are expected to be able to ‘use... syntax structures and vocabulary... correctly’, ‘negate sentences correctly’, ‘mutate correctly after prepositions and pronouns’, ‘spell correctly’ and ‘begin to craft their language and write accurately’ (DfES 2016, pp. 3-13), and the curriculum for primary school pupils learning Welsh as a second language mentions ‘accurate use of a variety of vocabulary, phrases, questions [and] sentence patterns’ (DfES 2015, p. 11).
A corpus-based analysis of educational curricula in the Arab States against European curricula and UNESCO benchmarks: Lexical networks and cognitive function of ICT terminology
Abeer Naser Eddine (University of Grenoble, France)

This research examines textual corpora of ICT terminology and concepts in the national educational curricula in the Arab world (referred to as Regional Corpus) versus European curricula (referred to as International Corpus). This contrastive approach allows us to outline extensive nomenclature of ICT in education in use over the last decade owing to the absence of standard definitions and indicators on the nature of ICTs integrated into and implemented in schools, despite efforts for standardisation. We focus on the curricula of ten Arab States that have policies, plans, and provisions to integrate ICT in education in general, and in First and Foreign Language Learning in particular, namely at the lower and upper secondary level of education in Lebanon, Jordan, Egypt, Qatar, Morocco, Palestine, Tunisia, Oman, United Arab Emirates, and Bahrain. By way of comparison, we observe their counterparts in European countries (including the UK). We also observe denotation-based definitions of ICT in education advocated by the UNESCO with the aim of developing a lexical network as well as highlighting the cognitive function of each.

To answer this question Is the use and attitude to ICT uniform?, we examine recurrent ICT terms in the Regional Corpus and the International Corpus. The frequency of a given term is defined as the sum of the frequencies of its morphological variants. This is to say, “competence” and “competencies” are referred to as two occurrences of the same term. As the patterns of word frequency in these two curricula do not entirely correspond, a “real” term (i.e. a word, a collocate, a word cluster or a multiword term) is considered familiar if it (co-)occurs ≥5 times in either curricula, with “at least one (co-)occurrence” (Leech et al. 2001) in one curriculum corresponding to one country. New terms familiar in either curricula (i.e. either corpora) are also extracted to emphasize the disparity of ICT terms in use. Besides, apparent incongruity between form and meaning between both curricula is included for contrastive purposes. Case-sensitivity is not taken into account. The citations are presented in a KWIC concordance format. Terms are classified by their thematic specificity, i.e. frequency in a given domain. They are extracted taking into consideration the notions of high unithood and high termhood as defined by Callon et al. (1986).

The emergence of word clusters like digitally supportive student, digital native, digital immigrant, digital confidence, digital humanities, digital competence, digital skills, digital literacy skills, etc. plus others like computer literacy, ICT-assisted instruction, computer-assisted instruction are not considered variants of the same concept. On the other hand, computing-related terms like programming knowledge, coding knowledge, robotics, robotic knowledge, etc. – more familiar in the International Corpus – are also considered thematically related to ICT. The domain ICT and its subdomains are outlined with the aim of creating lexical networks, and this includes:
To measure the compatibility degree of ICT terms in the two curricula, we first observe their frequency then we determine the semantic closeness/distance between clusters (Landauer et al. 1998). Second, we compare and contrast both curricula with UNESCOIBE Glossary of What Makes Quality Curriculum (2016) and Glossary of Curriculum Terminology (2013) as a “working reference” against which ICT capability can be assessed, knowing that the latter “is not (intended) to establish standard universally applicable definitions”. In this respect, in order to monitor polysemy, we consider higher-order thinking skills, a concept introduced in Bloom’s Taxonomy (1956) and later expanded to include the cognitive processes of analyzing, comparing, evaluating and creating (Anderson and Krathwohl 2001). In terms of the cognitive processes, we particularly investigate the extent to which ICT knowledge and skills included in the curricula incite higher-order and inventive thinking essential to enhancing students’ learning process (Kong et al. 2014).

We base our taxonomy of the cognitive domain on the theory of cognitive abilities proposed by Carroll (1993), structured into three-strata with different levels of breadth, moving from general to specific: Stratum three refers to general intelligence and it includes two sub-strata. Stratum two suggests eight cognitive abilities, and stratum one suggests a group of sixty-nine specific cognitive abilities underlying eight primary factors like language, reasoning, memory and learning, visual perception, auditory perception, idea production, cognitive speed, and knowledge and achievement. These abilities allow for further grouping as this model has been designed with this in mind, that is to say, to increase the validity and reliability of assessing (the subject’s) cognitive processes.

As a conclusion, based on the research findings, a more precise definition of ICT in its more precise or narrower sense is delineated, considering the differences in cognitive abilities required in the curricula of the Regional Corpus and the International Corpus.

References


developing 21st century skills: Critical research issues and policy implications. 
*Educational Technology and Society*, 17(1), 70-78.
http://unesdoc.unesco.org/images/0024/002439/243975e.pdf
Descriptive Translation Studies, and more specifically the sub-area of Corpus-based Translation Studies, have focused on the style of professional and literary translators, and addressed the translator’s discursive presence in the translated text as a result. These studies have also examined large translation corpora in order to identify possible features of translated texts. Nevertheless, studies considering the association between descriptive translation studies and stylistics are still few in number and scope. In addition, they have English language as its primary focus. The relationship established between style and conventional expressions in texts translated into Brazilian Portuguese accounts, in turn, for even more restricted results. This research draws on Corpus-based Translation Studies (BAKER, 2000; SALDANHA, 2011; WALDER, 2013) and sets out to investigate stylistic traits of a literary translator from the perspective of conventionality, creativity and shifts in translation. It specifically examines patterns of linguistic choices made by a translator in Brazilian Portuguese that could be found both in his work as a translator and as an author, and the consequences of these choices for the recreation of meaning in the translated texts. Shifts in translation in the context of previously selected conventionality/creativity items were also investigated in order to obtain more information about the translator’s linguistic preferences (PEKKANEN, 2010; BLAUTH 2015). The approach to the study of style in translation draws mainly on what was postulated by Munday (2008), Saldanha (2011) and Baker (2000). Saldanha (2011) takes into consideration the pioneering work of Baker (2000) about the style of translators in a comparable corpus, and proposes a combined approach to the investigation of style – the style of the translator as well as the style of the translated text using a combination of comparable and parallel corpora to differentiate translator’s choices from author’s choices. Baker (2007) introduces the study of conventionality associated with translator’s style and – to some extent – translator’s creativity, working with a comparable corpus. Munday (2008) investigates style from a different perspective than Corpus-based Translation Studies and identifies a certain degree of standardization regarding the voice of different authors when translated by the same translator. His results allowed him to make a connection between conventionality and creativity in translated texts. Magalhães and Barcellos (2015) focused on the investigation of translator’s style in parallel corpora without specifically addressing the study of conventionality in translated texts or its relationship to translator’s creativity. Therefore, the present research makes an attempt to fill in some of the gaps above mentioned through the investigation of conventionality as related to creativity and translation style in a corpus including translated texts and non-translated texts from the same translator/author. Thus, three corpora were compiled: 1) a corpus of translated texts written in Brazilian Portuguese by one of the current most prominent Brazilian literary translators, Paulo Henriques Britto, 2) a corpus of non-translated texts written
in Brazilian Portuguese by Britto, and 3) a corpus of short stories written in American English by the authors Philip Roth, John Updike, and Jhumpa Lahiri that, with the first corpus, translated texts by Britto, composed a parallel corpus. Two other corpora (COMPARA and ESTRA) were used as control corpora for frequency reference regarding the use of conventional expressions in Brazilian Portuguese. First, statistical data were obtained using the software WordSmith Tools © 6.0 (SCOTT, 2012), and elements related to conventionality in Brazilian Portuguese were analyzed at the various orders (morpheme, word, group, and clause). Second, shifts in the translated texts were investigated (CATFORD, [1965] 1978). The research methodology included compilation, preparation, alignment and tagging the texts for later analysis with WordSmith Tools © 6.0. The identification of patterns in the translated texts, attributed to the translator’s style and not to the linguistic constraints of the American English/Brazilian Portuguese pair, takes on board mainly what was postulated by Munday (2008), Saldanha (2011) and Baker (1999, 2000, 2007). The results indicated that Britto made a set of choices to some extent distinct for each translated text, under the influence of the style of source texts. In general, the linguistic choices made by Britto regarding the use of conventional expressions increased the degree of colloquialism in the translated texts when compared to their respective source texts. In addition, the set of choices identified in Britto’s non-translated texts presented similarities with the set of choices identified in his translated texts, in particular with the ones in Philip Roth’s work. The most frequent shift in translation was addition (an amplification subcategory). These instances of addition were not directly related to explicitation. They were, on the other hand, related to a preference from the translator to use conventional expressions in translated texts, even when there was no clear motivation for this in the source texts. Britto also made use of sanitization, erasing some cultural references from the source texts. Nevertheless, the translator’s creativity consistently outweighed the use of sanitization, corroborating the results obtained by Munday (2008) and refuting, to some extent, the ones obtained by Baker (1999, 2000).

References


Sarmento, & T. B. Sardinha (Eds.), *Pesquisas e perspectivas em Linguística de Corpus* (pp. 441-468). Campinas, SP: Mercado das Letras.


In 2014, 20 years after the last incidents of political terrorism marking Italy, Antimafia Judge Di Matteo received a threatening letter signed by the mysterious terrorist organisation Falange Armata. This name authored more than 1500 letters and phone calls from 1990 to 1994. Falange Armata alleged to be a terrorist organisation with a right-wing orientation whose main purpose was initially to call for better treatment of political prisoners by using threats of a future danger or, more frequently, claiming responsibility for recent episodes of crime and murders. Its targets were firstly prison operators and directors in cities across Italy. Later the focus switched to politicians, judges and journalists. Threats had been spread using media such as prison switchboards, newspapers, TV offices and press agencies. Investigations conducted by the head of police and the antiterrorism commission in 1994 had brought to trial only a single prison psychologist, Carmelo Scalone, who was later found not guilty.

This study, by examining the keywords, discourse and stance markers of the anonymous texts of Falange Armata through the corpus included in a legal report, collected by an unknown police officer under request of the Antimafia department, aims to analyse the stylistic and discursive constructions of this communication. The focus is on the analysis of key lexical and grammar features characterising the authors’ potential cultural background and their trends over time. The outcomes will help detect and profile the authorial style of this anonymous terrorist communication and to account for variation over a 4-year period.

Despite the lack of research on linguistic strategies and authorship profiling of forensic text types, specifically in the Italian context, it’s still possible to rely on the rising research framework coming from corpus based studies on genre and style (Biber, 1998, 2006), and discourse analysis on terrorist communication (Benedetti, 2002; Matusitz, 2012) and threats (Gales, 2011).

In the context of the only legal trial concerning Falange Armata, the investigative Antimafia department asked an unknown police officer to provide a report, locating and analysing a corpus collected by investigators under the name of the organisation. This report, which is the source of the study, divides the texts in 2 groups, consisting in the two ‘moments’ of the organisation, the “prison phase”, when the channel and targets are prison switchboards and prison operators and the “political phase”, when the channels are newspapers and press agencies and the targets are politicians and journalists. The report specifically talks about mythomaniacs, with a low educational level, without a purpose or a precise strategy, mocking more famous organisations such as Red Brigades and Ordine Nuovo. The author of the report also doesn't show any intention to provide a method that could help understanding how these statements are backed up. After many years, within the ongoing trial on a supposed negotiation between organised crime and government, some witnesses started releasing statements regarding the involvement of highly regarded members of the State within the organisation, with the purpose of promoting a climate of political destabilisation, the so-called strategy of tension. This led the judges in charge
of the trial to reopen the case to find out more regarding the authorship of these texts, which had remained controversial and mysterious ever since.

The attorney general of Palermo asked the author, in quality of linguistic consultant, to reanalyse the available data, especially in consideration of the outcomes of that report, which potentially affected the results of the trial and had an impact on the level of attention dedicated to the organisation in the 20 years before.

The research questions arising from this case study concern 2 aspects of these communications: How can we better profile the authorial style of Falange Armata texts using corpus based techniques? What information can linguistic data give us about the authorship’s cultural background?

The study combines the use of classic corpus linguistic tools with manual, computer-aided annotation and discourse analysis techniques, with a focus on stylistic preferences and lexical and grammar markers of stance. It aims to provide evidence that may confirm or disprove the anecdotal observations previously made on the nature of communication of Falange Armata and their authors. At the same time, it allows to experiment on the applicability of corpus based methods to Italian organised crime data for authorship profiling.

The corpus has been compiled by selecting and digitising 356 texts, totalling around 33,000 words out of the 519 of the report, excluding repeated messages and fragments. The decision to analyse such a small corpus was deliberate, as it allows for the combining of quantitative and qualitative corpus analytical techniques. It is partly speech tagged and lemmatized using TextPro (Pianta, Zanoli et al. 2008) and indexed with the Corpus WorkBench. Metadata encoded with the corpus include situational characteristics such as date, time, recipient details, target, delivery type (phonecall, letter, mixed), topic, text aims/moves.

To analyse the features characterising the communicative purpose of the texts, 4 different moves or genre markers have been identified and manually compiled and their occurrence and distribution have been calculated: threat, COR (claim of responsibility) communiqué and id (identification/signature). This approach to the texts, influenced by previous studies of move analysis and genres, was used to detect, on a discourse level, what the most predominant purposes of the message of Falange were.

Subsequently, a corpus lexical analysis of the texts was carried by using a combination of tools: AntConc to create a wordlist of the most frequent words in the corpus, Wordsmith to retrieve text by text information on lexical density and distribution, and Sketch Engine to extract a keyword list, detecting unusually frequent words and POS in the texts against the Ten Ten reference corpus Italian Web 2010 (POS tagged). Words and multiwords have been classified according to topics, keyness score and as register markers, i.e. specific to certain varieties of Italian language. This analysis allows to understand what the most salient lexical and grammar features characterising the texts are and how these features change over time.

Moving from the small to a larger-scale focus on the rhetorical and ideological construction of Falange Armata, the study carries on with the analysis of the authorial markers of stance (Biber, 1999). Some of the linguistic elements generally expressing or denoting involvement and interaction in general and specialised registers (Biber, 1988; Gale, 2012; Aull, 2014) were analysed. The features are: nouns, adjectives, pronouns, predicative verbs (agentive verbs, performative verbs, private verbs, public verbs), verbs introducing ‘that’ clauses and modal verbs. This approach appears to be
a suitable theoretical framework for interpreting author’s positioning towards his statements in threatening communication (Gale, 2010, 2012). Specifically, it accounts for the way in which the organisation represents itself, its targets and its strategies. The results have been integrated with a rhetorical analysis of idiomatic phrases and figures of speech in context.

The texts were divided, later in the analysis, into 2 subcorpora as they are distinguishable for situational characteristics (i.e. recipients, target, topics), stylistic choices and text metrics (such as TTR, word length and sentence length), to find how the two types compare.

The analysis offers an insight into ideology, style and variation in this terrorist discourse. The picture emerging is of a specific ideological construction expressed through the lexical and grammar choices, mostly coming from specialised registers, and highly polarised markers of stance, with distribution and conceptual domains changing over time.

The corpus based stylistic analysis of the keywords and collocations shows how the vocabulary features, registers and collocations included in the corpus address the authorial style profiling to a highly-educated military, legal, technical/bureaucracy background. A contrastive comparison between the 2 different subcorpora outlines a progressive trend into bigger lexical, density and variation in grammar usage and discourse conventions.

The results will be used by legal experts in ongoing trials on organised crime and terrorism, to frame, assess the style and authenticity of anonymous threatening and terrorist letters and aim to establish a corpus based tradition of studies along this line on organised crime and terrorist communication.

References


Community-specific language in online citizen science forums: a corpus-driven diachronic study
Claudia Viggiano (University of Portsmouth, UK)

Citizen Science can be defined as the collaboration between professional scientists and interested members of the public who together carry out scientific research (Socientize Project, 2013); it is considered to be a form of ‘crowdsourcing’ research. Online forms of citizen science take place on citizen science platforms, where volunteers usually tag, classify or annotate data. Currently the largest such citizen science platform, Zooniverse (www.zooniverse.org) is an umbrella website which hosts over 50 scientific projects, ranging from astronomy, e.g. Galaxy Zoo (www.galaxyzoo.org) to zoology, e.g. Chimp&See (www.chimpandsee.org). Each of the Zooniverse projects include a ‘Talk’ section where volunteers discuss science, ask scientists or moderators for help on the tasks they are carrying out, comment on their findings, or get to know one another. The present study as reported on in this paper is part of a wider project that uses corpus linguistics methods and tools to analyse the language of citizen science communities and, in particular, to investigate the relationship between online community building and short-term language change.

Many early and pioneering studies in sociolinguistics used a diachronic approach considering variation in language use according to social characteristics such as social class (Labov, 1966) or gender (Cheshire, 1982), doing so in the context of face-to-face communities, demonstrating, as in Trudgill (1974) how gender affects dialect in different social classes. Similarly, a recent trend in corpus linguistics (e.g. Aarts et al. 2013) is to study short-term diachronic change to investigate current change—changes in the language that have taken place over relatively short spans of time, using corpora. The present paper builds on these two approaches to diachronic change and group affiliation to analyse online citizen science communities.

Lave and Wenger (1991) defined a ‘community of practice’ as a group of people getting together to undertake certain tasks; Zooniverse communities, on the other hand, are an example of a virtual community of practice (Stewart, 2010). Due to their task-based nature and collaborative knowledge creation, they can be studied using the Community of Inquiry framework (Garrison, 2006), which identifies communities whose collective goal is based on empirical enquiry, and is often used to assess the success, and measure the learning outcomes, of pedagogic or e-learning platforms; for example Goertzen & Kristjansson (2007) found that the learning process of students on a distance learning programme was collaborative and deeply dependent on interpersonal engagement among participants. Accordingly, ‘social presence,’ the ability to participate personally and authentically in the community and to be perceived as salient and ‘real’ by others (Nichols 2009), is a part of the aforementioned community of inquiry framework. As such, there is an implication that the stronger the social presence and therefore the sense of community, the more productive said community is in achieving its goals. Lander (2015) identifies three different types of descriptors of social presence in the language used in online learning communities: (i) affective responses such as the
expression of feelings and emotions and the use of humour (I'm sorry I've been like this all day, I was up ALL night working on college work), (ii) cohesive responses such as greetings and inclusive pronouns (hello my beautiful galaxy friends), and (iii) interactive responses such as asking questions and quoting other members' posts (Allow me to pose a question in response to your question! Yes, answering a question with a question LOL!).

With such a framework in mind, the present work employs a 6 million-word corpus collected across 43 Zooniverse projects (Williams & Viggiano, 2016) to analyse short-term diachronic change and how language shifts therein are influenced by in-group dynamics; in particular, this paper looks at the introduction of new, community-specific terms: these are typically new lexical items or expressions that are introduced by one or more members, often from a semantic field related to the tasks undertaken by the community, and are used by members to signal their social presence and participation within the group. Hence, this paper looks at how the introduction and collaborative adoption of the terms by the community can be seen as cause and effect of social presence: on the one hand, the use of 'in-group' language is a marker of the participant’s strong social presence; on the other, a participant’s strong social presence can foster the creation of new ‘in-group’ language. The corpus downloaded from the Zooniverse website includes timestamp and poster information, which allow for the tracking of changes across time within the ‘Talk’ forums and to identify which forum members ‘lead’ change, and how the use of such terms can be, in itself, a marker of social presence.

Following a corpus-driven approach (Tognini-Bonelli, 2001), corpus query software Sketch Engine (Kilgarriff et al., 2014) is used to extract community-specific lexical items; in order to do so, a triangulation method of keyword and term analysis is employed: so as to determine lexical items unique to it, the Zooniverse corpus is set against three different sets of reference corpora—a general corpus, the New Model Corpus (https://www.sketchengine.co.uk/wp-content/uploads/New-Model-Corpus.pdf), an online corpus, enTenTen13, and an online scientific corpus, ScienceBlogs (both available on www.sketchengine.co.uk). Among the key items are terms with which members of Zooniverse self-identify as part of the community, thus creating group cohesion. For example, users self-identify as “zooites,” a term that is used to greet and address fellow volunteers:

(1) Ahhh.. good afternoon dear fellow zooites! :-* I've had a crazy few days and only just had a chance to catch up a bit.. (after snoozing on the sofa for a while!!) :)

Occasionally, the word is used to identify members as central or 'ordinary' “zooites,” therefore addressing an issue with expertise and status within the community:

(2) an ordinary zooite can notice something unusual in Radio Galaxy Zoo, and comment on it. [...] So you, other ordinary zooites, too can likely make stunning finds! :)

---

1 All three examples are drawn from my Zooniverse corpus (see below here).
Owing to the earlier mentioned time-stamping of posts as a form of metadata in the collected Zooniverse corpus, it is also possible to track the first few occurrences of the word, and how its use and meaning were collaboratively established with time and through discourse, displaying a strong interpersonal environment whereby meaning can be explicitly renegotiated:

(3) why are we "Users"? Surely there's a better term? Perhaps "zooites"? or "members" (we have to register and sign in)?
(4) You're all zooites in my eyes, but that term might confuse newbies! And "members" sounds a bit like something you pay for

Tracking the emergence, adoption and retention of community-specific language over time is a step towards the understanding of the collaborative creation of meaning in virtual communities, an aspect of the interpersonal dimension of meaning with which discourse studies informed by Systemic Functional Linguistics are concerned. In this regard, it builds on what Bednarek (2010) calls the instantiation of meaning—that is, how meaning serves as a means of both production and reproduction. As will be demonstrated in the paper, different such novel terms are picked up, maintained and re-negotiated in different ways according to matters of social presence (e.g. the initiating participant’s familiarity to the community, measured by the frequency and regularity of their posts). The mutual relationship between social presence and the creation and use of community-specific language can therefore be seen as a strong component of the community of inquiry, adding a ‘creative’ dimension to the social presence framework which Lander (2015) and Goertz & Kristjansson (2007) have found to be a central factor in the success of online communities.

References


Speech rates and unfilled pauses in native and learner speech at the CEFR levels B2 and C1
Lan-fen Huang (Shih Chien University, Taiwan) and Tomáš Gráf (Charles University, Czech Republic)

Learner corpora have been used to study the interlanguage of learners from different first language (L1) backgrounds and to compare their speech with that of native speakers to explore, for example, the overuse, underuse and misuse of various linguistic phenomena. However, little research has been done exploring spoken language of learners with different proficiency levels, possibly because learner corpora have been compiled on the basis of learners’ L1s, whose second language proficiency levels are not attested and are generally simply either inferred from educational background or from rating a small sample (e.g. 10% of the learners in the first eleven sub-corpora in the Louvain International Database of Spoken English Interlanguage (LINDSEI) (Gilquin et al., 2010)). This involves one of the methodological issues learner corpus studies have to address in order to have broader implications and applications in the fields of Second Language Acquisition and language testing and assessment (Callies & Götz, 2015).

The spoken data in the Czech (n = 50) (Gráf, 2015) and Taiwanese (n = 50) (Huang, 2014) components of LINDSEI were assessed on the scales of Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001) by two trained raters. They attended a four-hour rater standardization training session offered by an Examiner Trainer and their rating task followed the rating procedure developed by the LINDSEI team at the Catholic University of Louvain (Gilquin et al., 2010). The post-hoc assessment (r = 0.889 on global assessment) results in four proficiency groups: C2 (n = 2), C1 (n =38), B2 (n = 51) and B1 (n = 9). While most of the Czech learners (n = 36) are at C1 level, the majority of Taiwanese learners (n = 39) are at B2.

The present study aims to investigate two fluency variables, namely speech rates and unfilled pauses in the excerpts of a picture description task carried out by 38 C1 learners and 51 B2 learners in the Czech and Taiwanese sub-corpora of LINDSEI. The native counterparts were 50 speakers in the LOCNEC corpus (De Cock, 2004), performing the same task. Audacity (2013) and WordSmith 6 (Scott, 2012) were used to measure speech duration, calculate tokens and retrace unfilled pauses to explore their frequency and placement.

Accepting the common view that speech rates and the use of unfilled pauses are key components of L2 fluency (Kormos & Dénes, 2004; Ellis & Barkhuizen, 2005; Skehan, 2009; Ahmadian, 2012; Bosker et al., 2012), this study investigates whether significant differences in that regard exist not only between learners and native speakers as is commonly explored (Götz, 2013; Gráf, 2015) but also between learners across two proficiency levels on the CEFR scales.

The mean speech rates of the learners of C1 and B2 levels are 142 (sd = 20) and 118 (sd = 22) words per minute respectively and 174 (sd = 34) for the native speakers. Similar to the comparison of LOCNEC and the German (Götz, 2013) and Czech (Gráf, 2015) components of LINDSEI, the results indicate that the difference between learners’ and native speakers’ speech rates is statistically significant (p <
The difference between the two proficiency levels also proves to be statistically significant ($p < 0.00001$). It appears that rise in proficiency between these two levels is accompanied by increase in speech rate even at these relatively high levels of proficiency.

The C1 and B2 learners have been found to overuse unfilled pauses, pausing 10.76 and 14.43 times per hundred words (phw) respectively as opposed to the native speakers’ rate of 7.15 pauses phw. To determine differences in the use of unfilled pauses, these are further categorized as pauses at clause boundaries, within clauses, and within constituents. Significant differences are found between the learners and native speakers ($G^2 = 220.53; p < 0.0001$), and also between C1 and B2 learners in the use of unfilled pauses within clauses ($G^2 = 70.63; p < 0.0001$) and within constituents ($G^2 = 50.51; p < 0.0001$) but not at clause boundaries ($G^2 = 0.09; p > 0.05$). In terms of the locations, unfilled pauses at clause boundaries are taken as natural stops and those within constituents are hesitation stops, which are considered as markers of disfluency. In this respect, the C1 learners, who produced 1.02 pauses within constituents phw on average, performed better than B2 learners, who averaged at 2.20 pauses phw. The occurrence of pauses within constituents is much rarer in native speech – 31 out of 50 of our native speakers do not pause within constituents, and the remaining 19 do so much less frequently, at 0.45 pauses phw.

The investigation of speech rates and unfilled pauses reveals that both can be used to distinguish native from learner speech as well as differentiating between the two adjacent proficiency levels. Both levels of learners pause more frequently than native speakers. This is in line with Pawley and Syder's (1983) argument that native speakers pause or slow down typically at or near clause boundaries and only rarely in the middle of clauses. The location of pauses would thus appear to be an aspect the learners might like to concentrate on if they desire to sound more fluent.

Whilst this study takes into consideration only two fluency variables, the results shed some light on fluency variation at C1 and B2 levels. With further analysis of other variables, such as filled pauses, repeats and mean length of runs, research studies of this nature will have the potential to be utilized in the teaching of oral skills and assessment of speech competence.

References

Ahmadian, M. J. (2012). The effects of guided careful online planning on complexity, accuracy and fluency in intermediate EFL learners' oral production: The case of English articles. Language Teaching Research, 16(1), 129-149.


Textbook English: a corpus-based analysis of language use in German and French EFL textbooks
Elen Le Foll (Osnabrück University, Germany)

Given that textbooks account for a considerable proportion of learners' foreign language input (Kim & Hall, 2002; Vellenga 2004), this PhD project aims to provide a linguistic description of the lexico-grammatical content of popular English textbooks used in secondary schools in France and Germany. This poster presents the methodology design and preliminary results.

Mindt’s pioneering textbook corpus analysis work (e.g. 1987, 1992, 1995) compared the frequencies, functions and co-occurrences of future expressions found in German EFL textbooks with those from a reference corpus of spoken British English. Since, a number of textbook corpora studies following the Mindtian approach – comparing textbook corpora with “authentic” native corpora – have emerged. So far, these have largely focussed on specific grammatical aspects, such as the use of the present perfect (Schlüter, 2002), progressive forms (Römmer, 2005), if-sentences (Römmer, 2004a; Gabrielatos, 2006), reported speech (Barbieri & Eckhardt, 2007) and constructions typical of spoken grammar (Cullen & Kuo, 2007). These studies reveal considerable discrepancies between “authentic and natural” reference corpora and the language presented in textbooks. However, to the author's best knowledge, no attempt has yet been made to provide a general lexico-grammatical description of textbook English, nor have EFL textbooks targeted at the same level, but designed for use in different countries been compared before.

Thus, this study’s first aim is to draw up a general description of the lexico-grammatical content of textbook English following the bottom-up corpus analysis approach advocated, among others, by Biber (1998), Biber & Gray (2015) and Carter & McCarthy (2006). The analysis will include both grammatical features explicitly taught in the examined textbooks such as the use of adjectives, adverbs and verb forms, as well as aspects that are almost exclusively presented in an implicit manner, such as the use of prepositions, collocations and colligations.

The textbook corpus consists of all the texts included in six full series of secondary school EFL textbooks which have been annotated to enable comparisons between text genres, learner level and learner L1. The native reference corpus is subdivided into a number of sub-corpora which reflect the different types of texts featured in textbooks (e.g. conversational language, informational texts and fiction). Comparing the lexico-grammatical content of the textbook corpus and its sub-corpora with the relevant native reference sub-corpora will enable quantitative and qualitative comparisons of the main differences between “the kind of synthetic English that pupils are often confronted with in the classroom” (Römmer, 2004, p. 185) and the reference corpus. Furthermore, the degree of authenticity of both the textbooks series and the curricula and pedagogical approaches applied in the countries of the textbooks investigated will be compared.

Preliminary results arise from a comparison of the frequency of occurrence of linguistic features typical of conversations in both the spoken textbook sub-corpus and the spoken British English reference sub-corpus. Major differences are illustrated with examples from both sub-corpora.

It is hoped that pointing to such substantial discrepancies between textbook English and naturally occurring English can contribute to improving the relevance and efficiency of the pedagogic texts presented in EFL textbooks.

References


A Case Study of the Characterizations of Spoken Academic English: The Adjective-Noun Combination
Fuying Lin (Free University of Berlin, Germany)

1. Introduction

"Academic English" constitutes a variant of English with properties unique enough to describe it as register (in theoretical linguistics) and to teach it with a specific case of English for Special Purposes (in applied linguistics) (cf. Brisk & Jeffries, 2008 and Bailey, 2012 for overviews and Benesch, 2001 for a critical appraisal). Academic English is largely treated as a written register with occasional exceptions, which often tends to be tiny portion of language materials; for example, spoken materials used in the New Academic Word list (NAWL 1.0, cf. Browne, C. et.al, 2013) is merely 1%, which is unlikely to have a noticeable effect on the word selection process. Although theoretical researchers are more inclusive of spoken language on average, researchers like Biber and Grey also equate Academic English with written academic prose in their investigation (Biber & Grey, 2016). Where spoken language is included, the definition of "Academic English" is often too broad, covering any type of classroom discourse, for instance, in the Corpus of English as Lingua Franca in Academic Settings (ELFA 2008) or in the Michigan Corpus of Spoken Academic English (MICASE) (Simpson, Briggs, Ovens, & Swales, 2002). And with the increasing popularity of Massive Open Online Courses (MOOCs), the visibility of the spoken contents in Academic English is worthy of more severe manifestation (see Baiely, 2012). This paper, seen as a pilot study of my dissertation, will narrowly aim at the investigation of the adjective-noun combination structure manipulated in Spoken Academic English.

2. Background

2.1. MOOCs and their significance

MOOC is an abbreviation of Massive Open Online Courses, a term that emerged in 2008 for a particular type of Internet-based long-distance learning format that has become a new extension of higher education in recent years (Fini, 2009; Kennedy, 2014). The year 2012 was even identified as the year of MOOC due to the enormous number of platforms, organizations and institutes joining the MOOC movement with offerings of their own (Pappano, 2012). There are at least 26 known providers of MOOCs, three of which can be considered major providers—Coursera, Edx and Udacity (Ha, 2014).

MOOCs clearly present a novel learning opportunity, making Higher Education available to many potential students all over the world. Unlike previous forms of self-study and/or long-distance learning, however, the participation in MOOCs requires a thorough command of academic English, not just in its written but also in its spoken form. This issue has not received enough attention in the research on academic English in general.
2.2 Definition and significance of Spoken Academic English

Several scholars have tried to distinguish academic English from general English (Coxhead and Nation, 2001; Schmitt, 2000; Xue and Nation, 1984). Academic English, different from everyday spoken English, has its own professional vocabulary (not limited to technical terminology), and is grammatically distinct with its complex, informationally-dense structures (Biber, Finegan, Johansson, Conrad, & Leech, 1999). There is an implicit equation of academic English with written English, even in recent research that challenge some of the stereotypical views of this register, such as Biber & Gray (2016). And although a few corpora that could be called Spoken Academic English have been constructed, such as 1.8-million-word Michigan Corpus of Spoken Academic English (MICASE) corpus (Simpson et al., 2002), the British Academic Spoken Corpus (BASE) (Nesi & Thompson, 2001), the Corpus of English as Lingua Franca in Academic Setting (ELFA 2008), and the 2-million-word Hong Kong Corpus of Spoken English (HKCSE)(Warren & Chen, 2004), they collected language data from various academic settings like classrooms, meetings, seminars, conferences and job interviews.

For the objective of my research, I will define Spoken Academic English more narrowly as the spoken language used by academics in talking about their subject in academic settings. This includes talks by and discussions amongst researchers at academic conferences as well as lectures by instructors in (physical or virtual) university classrooms, but excludes academics using language in non-academic university settings (such as faculty meetings), language dealing with organizational aspects of teaching, and classroom language by non-academics (i.e. students). The latter may be considered a learner variety of Spoken Academic English, worthy of study in its own right. The language of MOOCs, which is the focus of this research, clearly falls under this definition.

The Spoken Academic English found in the context of MOOCs is worthwhile not just in terms of characterizing it as an (emerging) register, but also with a view to improve ESP teaching and self-study resources. Some scholars have argued that an identifiable subgenre within Spoken Academic English from MOOCs justifies "EMP" (English for MOOC purposes, see Anthony, 2015).

3. Methodology

Because the current corpora of Spoken Academic English are either too small (such as BASE with 1.6 million tokens)(Thompson & Nesi, 2001), or proprietary and inaccessible (such as PICA, Ackermann et al., 2010), to achieve the aims of my research, I have specifically constructed a corpus, the MOOC corpus, consisting of 93 lecture transcripts from MOOCs, amounting to 8,716,104 tokens, in five broad subjects represented offered by the major MOOC providers (Coursera, EdX, Udacity and Futurelearn), as shown in the table 1 below.

The subjects are those widely agreed upon in studies of Academic English, represented in such large corpus as the British Academic Written English Corpus(BAWE)(Garnder & Nesi, 2012, 2013).

In order to identify the unique properties of Spoken Academic English, a corpus of general spoken English (the spoken part of COCA, the Corpus of Contemporary American English)(Davies, 2010) will be treated as the reference corpora.
Table 1. Overview of the MOOC spoken academic corpus

<table>
<thead>
<tr>
<th>category</th>
<th>subcategory</th>
<th>course number</th>
<th>Total course number</th>
<th>Word count</th>
<th>Total word count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arts and humanities</td>
<td>language &amp; literature</td>
<td>5</td>
<td>25</td>
<td>252,219</td>
<td>1,855,055</td>
</tr>
<tr>
<td></td>
<td>history</td>
<td>5</td>
<td></td>
<td>610,621</td>
<td></td>
</tr>
<tr>
<td></td>
<td>art, architecture &amp; design</td>
<td>5</td>
<td></td>
<td>312,465</td>
<td></td>
</tr>
<tr>
<td></td>
<td>music and design</td>
<td>5</td>
<td></td>
<td>317,788</td>
<td></td>
</tr>
<tr>
<td></td>
<td>philosophy</td>
<td>5</td>
<td></td>
<td>361,962</td>
<td></td>
</tr>
<tr>
<td>Social science</td>
<td>sociology</td>
<td>4</td>
<td>20</td>
<td>261,319</td>
<td>1,626,268</td>
</tr>
<tr>
<td></td>
<td>political science</td>
<td>4</td>
<td></td>
<td>328,699</td>
<td></td>
</tr>
<tr>
<td></td>
<td>education</td>
<td>4</td>
<td></td>
<td>211,313</td>
<td></td>
</tr>
<tr>
<td></td>
<td>economics &amp; finance</td>
<td>4</td>
<td></td>
<td>413,999</td>
<td></td>
</tr>
<tr>
<td></td>
<td>law</td>
<td>4</td>
<td></td>
<td>410,938</td>
<td></td>
</tr>
<tr>
<td>Life science</td>
<td>biology</td>
<td>4</td>
<td>16</td>
<td>468,818</td>
<td>1,416,234</td>
</tr>
<tr>
<td></td>
<td>Health science</td>
<td>4</td>
<td></td>
<td>294,882</td>
<td></td>
</tr>
<tr>
<td></td>
<td>food &amp; nutrition</td>
<td>4</td>
<td></td>
<td>169,415</td>
<td></td>
</tr>
<tr>
<td></td>
<td>psychology</td>
<td>4</td>
<td></td>
<td>483,119</td>
<td></td>
</tr>
<tr>
<td>Physical science</td>
<td>physics, astronomy</td>
<td>4</td>
<td>16</td>
<td>851,967</td>
<td>1,822,527</td>
</tr>
<tr>
<td></td>
<td>chemistry</td>
<td>4</td>
<td></td>
<td>332,931</td>
<td></td>
</tr>
<tr>
<td></td>
<td>engineering</td>
<td>4</td>
<td></td>
<td>213,981</td>
<td></td>
</tr>
<tr>
<td></td>
<td>energy &amp; earth science</td>
<td>4</td>
<td></td>
<td>428,648</td>
<td></td>
</tr>
<tr>
<td>Information, computer science &amp;</td>
<td>computer science</td>
<td>4</td>
<td>16</td>
<td>621,880</td>
<td>1,996,020</td>
</tr>
<tr>
<td>mathematics</td>
<td>robotics &amp; electronics</td>
<td>4</td>
<td></td>
<td>347,827</td>
<td></td>
</tr>
<tr>
<td></td>
<td>mathematics</td>
<td>4</td>
<td></td>
<td>411,320</td>
<td></td>
</tr>
<tr>
<td></td>
<td>data science</td>
<td>4</td>
<td></td>
<td>615,533</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>93</td>
<td></td>
<td>8,716,104</td>
</tr>
</tbody>
</table>

4. Results & Conclusion

Academic English is assumed to be characterized both by domain-specific vocabulary and by vocabulary also found in colloquial language but used with specific technical meanings in academic settings. Terminological vocabulary may consist of single words, but more often it consists of multi-word expressions of sequences like Adjective-Noun or Noun-Preposition-Noun (Justeson & Katz, 1995). Comparing the frequency of, in this paper, Adjective-Noun combinations in the MOOC corpus against their frequency in the spoken part of the COCA and extracting combinations that occur significantly more frequently in the former than in the latter should thus allow us to identify academic terminology.

Table 2 shows the top 15 adjective-noun combinations that are significantly
more frequent in the MOOC data than in the spoken part of the COCA with their observed and expected frequencies, the log-likelihood value and the information whether they occur in both corpora or only in the MOOC data.

**Table 2.** Ad-N combinations significantly more frequent in the MOOC corpus compared to COCA

<table>
<thead>
<tr>
<th>WORD</th>
<th>MOOC data Observe Expected</th>
<th>COCA spoken data Observed Expected</th>
<th>LL</th>
<th>Shared</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 random variable</td>
<td>1419 320.8 0</td>
<td>1098.2 4222.74 No</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 random variables</td>
<td>834 188.6 0</td>
<td>645.4 2481.05 No</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 special theory</td>
<td>722 163.9 3</td>
<td>561.1 2110.35 Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 fair use</td>
<td>465 106.7 7</td>
<td>365.3 1313.77 Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 next video</td>
<td>467 109.2 16</td>
<td>373.8 1256.68 Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 next lecture</td>
<td>348 78.9 1</td>
<td>270.1 1021.78 Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 differential</td>
<td>343 77.8 1</td>
<td>266.2 1006.94 Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 gradient descent</td>
<td>338 76.4 0</td>
<td>261.6 1005.23 No</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9 absolute value</td>
<td>315 72.4 5</td>
<td>247.6 887.87 Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 initial state</td>
<td>298 67.8 2</td>
<td>232.2 863.24 Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11 solar system</td>
<td>529 175.9 249</td>
<td>602.1 725.50 Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12 white dwarf</td>
<td>240 54.5 1</td>
<td>186.5 701.28 Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13 kinetic energy</td>
<td>255 60.1 11</td>
<td>205.9 672.36 Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14 straight line</td>
<td>371 197.6 105</td>
<td>368.4 654.86 Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15 conditional</td>
<td>219 49.0 0</td>
<td>169.5 651.27 No</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The approach clearly yields mostly scientific terminology, either in a narrow (domain-specific) sense (like *random variable(s)*, *differential equation*, *gradient descent*, *kinetic energy*) or in a general sense (*fair use*, *absolute value*, *initial state*, *solar system*). The former tend to be restricted to the MOOC data, while the latter tend to occur in the general language too, but have a special technical meaning in academic English.

**References**


Benesch, S. (2001). *Critical English for academic purposes: theory, politics, and*
Fini, A. (2009). The technological dimension of a massive open online course: The Case of the CCK08 Course Tools. The International Review of Research in Open and Distance Learning 10 (5). doi: http://dx.doi.org/10.19173/irrodl.v10i5.643
Growth in Grammar: A multidimensional analysis of student writing between 5 and 16
Philip Durrant and Mark Brenchley (University of Exeter, UK)

This poster will report on an in-progress project investigating the development of written language in school children in England. The study aims to provide a more thorough understanding of linguistic development in writing and to inform national and international curriculum policies on the teaching of English. It will also generate an updatable corpus of linguistically-annotated, educationally-authentic student writing, which will be made accessible to researchers and teachers.

We will address the following research questions:

1. What combinations of linguistic features distinguish school students’ writing across:
   a. different year groups?
   b. different levels of writing attainment?
   c. different text genres students are expected to write?

2. How do these combinations compare with those found in adult writing?

To answer these questions, we are building a new corpus of children’s writing, sampled from schools across England. To capture the range of variation found in children’s writing, the contents of the corpus will be balanced across year groups, levels of attainment, academic disciplines, text genres, geographic locations, student gender and student socio-economic status.

We will then use a combination of existing and custom-built software to identify and count linguistic features of potential interest (as identified by a systematic review of the research literature) in these texts. These counts will form the basis of a series of multi-dimensional analyses to determine the key dimensions of variation across ages, genres, and levels of attainment.

On completion of the project, we will be making our corpus available for other researchers and teachers to use. We will also be creating an online interface which will enable users to explore the corpus for themselves.

This poster will describe the design of the corpus and methods of analysis and report on progress to date.
Evaluating the effect of data-driven learning (DDL) on the acquisition of academic collocations by advanced Chinese learners of English
Tanjun Liu (Lancaster University, UK)

This study aims to explore the effect of data-driven learning (DDL) on the acquisition of academic collocations by advanced Chinese English as a Foreign Language (EFL) learners in a Chinese university. Collocations, prefabricated multi-word combinations, are considered to be a crucial component of language competence which indicates the central role they should play in language teaching and learning. However, collocations remain a challenge to L2 learners at different proficiency levels, and particularly a difficulty to Chinese learners of English (e.g. Fan, 2009; Granger & Bestgen, 2014). From a pedagogical perspective, collocations have so far attracted only limited attention in language teaching in the Chinese language teaching classroom. This study, therefore, focuses on the effectiveness of the teaching of academic collocations to advanced Chinese learners of English, using a specific pedagogical approach, the corpus-based data-driven learning approach (DDL). DDL has been argued to offer an effective teaching method in language learning. However, so far the large-scale, quantitative studies carefully evaluating the effectiveness and assessing the benefits of DDL in the acquisition of academic collocations are limited in number when compared to a different method of teaching of collocations (Bouton, 2010; Chambers, 2005).

This study, therefore, contrasts DDL with another method of teaching of collocations (using of online dictionaries), with regards to the breadth and depth of the knowledge of academic collocations. The study uses data from 120 Chinese students of English from a Chinese university and employs a quasi-experimental method, using a pre-test-and-post-test (including delayed test) control-group research design to compare the achievement of the use of DDL and online dictionary in teaching academic collocations to the Chinese EFL learners. The experimental group uses #Lancsbox (Brezina, McEnery & Wattam, 2015), an innovative and user-friendly corpus tool. By comparison, the control group uses the online version of the Oxford Collocations Dictionary. The results are analysed for the differences in collocation gains within and between the two groups. Those quantitative data are supported by findings from semi-structured interviews exploring learners’ attitudes towards DDL and linking their attitudes with the test results. The findings contribute to our understanding of the effectiveness of DDL for teaching academic collocations and suggest that the incorporation of technology into language learning can enhance collocation knowledge.

References

Changing Discourse-Semantics of Risk - A historical analysis of US newspapers after World War II utilizing corpus linguistics research instruments

Jens O Zinn (Lancaster University, UK) and Daniel McDonald (University of Tübingen, Germany)

Since the 1980s and 1990s the notion of risk has become increasingly influential in societal discourses and scholarly debate (Skolbekken, 1995). From early work on risk and culture (Douglas, 1966, 1992) to the risk society thesis (Beck, 1992, 2009; Giddens, 2002), from governmentality theorists working in the tradition of Foucault (Dean, 1999; O'Malley, 2012; Rose, 1999) to modern systems theory (Luhmann, 1993) all have built their work around the notion of risk and implicitly or explicitly refer to linguistic changes. Though this body of literature offers different explanations for the shift towards risk and its connection to social change, to date there have been few attempts to empirically examine their relative ability to explain this change. At the same time, linguists have approached risk from frame-semantic and corpus linguistic perspectives, building a conceptualisation of the kinds of participants and circumstances that co-occur with represented risk events. These approaches, however, have not attempted to identify longitudinal change, or to connect their work to sociological accounts of risk.

This presentation reports from an ongoing research initiative (compare Zinn & McDonald 2015, 2016; Zinn 2010) which aims to advance understanding of the historical changes surrounding risk, and to develop the corpus/computational tools for extracting useful information from parsed and metadata-rich corpora. First, we provide a general overview of our project, describing novel possibilities for sociological analysis afforded by increasing digitisation of news and increasingly fast and accurate tools for annotating and parsing natural language. Second, we introduce key claims made within sociology about the nature and evolution of the concept of risk. Variously, it has been argued that risk is increasing understood as negative, that risk is increasingly scientised, and that institutional resources are increasingly devoted to the identification and management of particular kinds of risk. At the same time, we are mindful of insights from media studies, which emphasize the fact that news coverage is significantly influenced by factors such as the political stance and style of reporting.

Our methodology involved the creation of a large, grammatically annotated, metadata rich text corpus of digitized newspaper archives. In order to control for newspaper specific factors such as a particular stance, thematic orientation, style of journalism and regional/national focus in reporting, we included a number of different newspapers: the New York Times, the Wall Street Journal, The Tampa Bay Times, USA Today, Chicago Tribune and the Washington Post. We included all articles where a risk token has been used at least once in the years 1987 to 2014, preserving metadata concerning topic, date, author, publication and publication section where possible.

We parsed all articles using Stanford CoreNLP (Manning et al. 2014) and developed corpkit, a purpose-built tool for performing detailed analysis of parsed
corpora. Using the tool, we queried the constituency and dependency annotations, uncovering sites of change within the lexicogrammar of risk clauses. Concepts from systemic functional grammar (Halliday & Matthiessen, 2004) were used to relate Mood and Transitivity patterns of risk clauses to changes in the arguability and experiential semantics of risk. Finally, longitudinal changes in risk language are then mapped to key events, as well as broader social processes and changes.

The presentation will present evidence for systematic discourse-semantic changes across newspapers but also systematic differences between newspapers and social domains:

- A growing routinisation and institutionalisation of risk across all newspapers.
- Decreasing human agency in risk processes (e.g. *risk, take risk, run risk, put at risk*). There is also clear evidence for the discourse semantic of risk-taking being affected by newspaper specific factors in contrast to other risk processes.
- A clear increase of reporting which presents people and particular social groups as lacking control, especially regarding health issues.
- Risk reporting, in particular in the health sector, being driven by reference to scientific research supporting a rationalised approach to risk. However, expressions of control such as calculated risk are decreasing while expressions indicating the possibility of negative outcomes are increasing indicating a rise of a possibilistic notion of risk.
- A clear difference in the presentation of powerful risk-takers and relatively powerless at-risk groups. The difference between the groups is increasing over time. The powerful take more social risks while the powerless take much more substantial risks often related to illness, injury and death.

Discussion is organised around two main themes. First, we describe theoretical challenges and compromises inherent to a project involving constituency, dependency, frame-semantic and systemic functional linguistic theory, highlighting strengths and limitations of the various linguistic theories and grammars as ways of extracting and interpreting results from corpora. Second, we provide an account of the relationship between identified semantic change and sociological hypotheses about risk. Trends toward nominalisation, reduced arguability and increasing use of risk as a pre-head nominal modifier, for example, have only partial compatibility with Beck’s account of the Risk Society. Future research aims are then briefly presented.

**References**


Opinions on the style of The Lord of the Rings differ widely, ranging from scathing criticisms to the effect that “prose and verse are on the same level of professorial amateurishness” (Wilson, 1995, p. 59) to praise of Tolkien’s ordinary yet dignified diction (Kirk, 1977, Rosebury, 2003, p. 22).

The present study takes a more objective, corpus-stylistic approach to Tolkien’s masterpiece, asking whether it displays any stylistic features that are either significantly more common than in other texts or even distinctively its own. Among the characteristics considered are word frequency, key words, key collocations, key semantic areas and key n-grams.

Two separate methodologies were used for computing key words and key collocations, respectively. With regard to key words, a detailed comparison was undertaken between key words in 19th century fiction, late 20th century fiction and the Lord of the Rings, as compared to a ‘mixed’ reference corpus. The 19th century corpus was about 30 million words in size, comprising mainly canonical writers such as Jane Austen, Thomas Hardy, Charlotte and Emily Brontë, George Eliot, Elizabeth Gaskell, George Meredith, etc., with whose writings Tolkien was almost certainly acquainted. The post-war fiction corpus comprised around 160 million running words and was fairly evenly balanced between eight major genres (belles lettres, general fiction, science fiction, adventure, romance, mystery and detective, fantasy, and children’s fiction). Key words were computed using the Key words tool of WordSmith 5.0 (Scott, 2008). The tool settings were as follows: minimum frequency = 3, maximum key words = 7,500; p value = 0.000001; procedure = log likelihood, meaning that a word, or rather a word form, was considered to be key if a) it occurred at least three times in the text and b) turned out to be outstandingly frequent when compared to the reference corpus by means of Dunning’s (1993) log likelihood function. Following the findings of Scott (2009), which suggest that an optimum reference corpus should be both large and heterogeneous, a reference corpus of 94 million words was collected, comprising volumes II, III, IV and IX of the Cambridge History of English and American Literature (1 m), a wide range of academic texts from the National Academy Press (10 m), a general encyclopedia (10 m) and a medical encyclopedia (1 m), the Times on CD-ROM 1995 (30 m), the Cheddar Valley Gazette and the Brentwood Gazette 2006 (2 m), excerpts from Hansard (the verbatim report of proceedings in the House of Commons and the House of Lords) (35 m), and a selection of transcripts from American talk shows and news programs (5 m). In addition, Wmatrix (Rayson, 2008) was used to identify key semantic fields.

Key collocations were computed by means of a comparison between two lists, one containing collocations specific to Lord of the Rings by comparison with the imaginative section of the BNC, the other containing collocations specific to Lord of the Rings by comparison with a 30-million word corpus of nineteenth-century novels. First, the individual lists were compiled using the Sketch Engine’s word list function (‘collocation’ setting). Next, collocations that appeared in either of the two reference
corpora and Lord of the Rings were eliminated from the lists, so that only collocations specific to Lord of the Rings remained.

A number of interesting findings emerge from the key word analysis. Most importantly perhaps, it is evident that the impression of archaicty which any reader will experience on reading Lord of the Rings is partly due to three simple lexical causes: the ‘overuse’ of words borrowed from nineteenth-century fiction, the avoidance of words associated with the modern world, and the comparatively dense use of new coinages, unusual grammatical patterns, rare or obsolescent words.

What is particularly noteworthy with respect to key words is the dense presence of nineteenth-century words from four lexical sets: (1) landscape description; (2) position or direction; (3) warfare and the struggle between good and evil; (4) other descriptive verbs and adjectives. Among the highest scoring 200 key words shared with nineteenth-century novels we find a large number of terms describing features of the landscape as well as position or direction, such as mountain(s), hills, hilltop, mountain-side, topmost, bough, hiding-place; road, river, cross-roads, passage, winding; above, westward, northward, southward, near, under, whither, thither, whence, upwards, yonder, afar; journey (v), hewn, descend. While it might be argued that these are reasonably to-be-expected characteristics of a novel revolving around a prolonged journey, it is still significant that so many of the keywords used to describe its settings were obsolete or obsolescent at the time of writing. Another interesting finding is that Tolkien borrows a host of words from his translation of Beowulf (e.g. valour, valiant, corslet), stretching them creatively to cover a wider range of uses than was possible in Old or Middle English (e.g. Branched lightning smote down upon the eastward hills.)

As for the modern key words found in Lord of the Rings, it is worth noting that these are usually part of the core vocabulary of present-day English, such as they, big, wake, wall, unfriendly, tired or smoke. However, Tolkien tends to use many of these words in nineteenth-century patterns. One example among many is smoke, which is commonly preceded by the indefinite article in Lord of the Rings (e.g. ... a black smoke swirled in the air.), a use that is notably infrequent in modern fiction. One area in which Tolkien tends to follow modernist tendencies is his abundant use of descriptive verbs (i.e. verbs which apart from denoting an event contain an additional semantic feature that assumes the function of a manner adverb, (cf. Snell-Hornby, 1983, p. 25) and, to a lesser extent, nouns (e.g. haze, stench). Among the key verbs shared with modern fiction are the following: clutch, cower, crack, crawl, curl, flicker, frown, gape, gasp, gaze, gleam, glimmer, glitter, growl, hiss, huddle, loom, paw, peer, pile, plod, reek, shroud, slant, stumble, stow, stride, stumble, twist.

Turning now to those key words that are exclusive to Lord of the Rings (i.e. non-key in 19th and 20th century fiction) and hence particularly distinctive of Tolkien’s style, we find that these, too, fall into a number of relatively distinct semantic categories: warfare (e.g. armies, armouries, besieged, destroyed), features of the natural environment (e.g. branches, coasts, dike, downs, fen, fern, dingle, glades, land, starless, starlight, thickets; hewed, hewn, encircling), artefacts made by the inhabitants of Middle Earth (fastness, cities, boats, horn, sheath, shield), people and animals (dwarves, herdsmen, lore-masters, gaffer, hornblower, horse-men, riders, wolf-riders; dragon, steed), time expressions (season, yesteryear, yuledays) as well as certain prepositions and adverbs (along, among, outwards). As
can be seen from the above listings, most of these words are comparatively rare in present-day English and, although not archaic, work together to create an atmosphere of medievalism or otherworldliness.

As is to be expected given the logic of numbers, many of the significant collocations in Lord of the Rings are based on key words such as road, altogether, ancient or mighty, to give a few randomly chosen examples. A cursory look at the collocational range of certain key words suggests that Tolkien shows a marked preference for the most homely collocations available. However, Tolkien also uses everyday adjectives in what would have been considered unusual collocations even in his day. Great, for example, is most commonly associated with three meaning groups: abstract nouns such as peril, deed, force, power, strength; concrete objects such as trees, stones, halls, horns, gates; natural phenomena such as clouds, storms, shadow, wind or smoke. These collocations are derived from 19th century fiction, where they occur much less frequently, though.

While we have already noted Tolkien’s preference for geographical terms, darkness and light, location and direction as well as colour adjectives, an analysis focussing on key semantic domains, apart from providing further detail about the domains just mentioned, highlights other domains which key word analysis misses entirely. One such domain is ‘plants’, where certain affinities are found with nineteenth-century fiction (e.g. reeds), but the domain is in fact much larger than suggested by the key word analysis, including the following items:

tree: elm-tree, alder-tree, fir-tree, pine, beech, acorn, willow; lobelia, nightshade, hemlock, rowan, ivy, hyacinth, marigold, hawthorn, fern, holly, shrub, sward, etc.

Other major sources of descriptivity include ‘sailing/swimming’, ‘shape’, ‘weight: heavy’, ‘speed: fast (slow)’, ‘entire/maximum’, ‘sound: loud’, ‘temperature: hot/on fire’, ‘long, tall and wide’, the last five domains lending additional weight to the hypothesis that Tolkien tends to resort to hyperbole. Semantic domains which are also prevalent in the realist novel (cf. Kullmann, 2009, p. 44) and partake in the creation of circumstantial realism include ‘general appearance and physical properties’ (e.g. hard, stony, bare, plain, bow, kneel, hewn, bold, hollow, blank, splendour, furry, weather-beaten) ‘geographical terms’ (e.g. land, forest, mountains, marshes, vale, wilderness, cave), and, to a minor extent, ‘farming and horticulture’ as well as ‘fixing and mending’.

In summation it can be said that, far from being ‘amateurish’, Tolkien’s prose effectively employs several centuries’ worth of linguistic developments, putting old words to new uses and vice versa. In a much less obvious sense than in the case of James Joyce or Virginia Woolf, he is a linguistic postmodernist.

References


Corpus-assisted editing for doctoral students: Do-it-yourself corpora for self-correction and learning
Maggie Charles (Oxford University, UK)

From the early days of data-driven learning, one impetus for the pedagogical use of corpora arose from the need to deal with errors or infelicities in students’ written work (Johns, 2002). Since then, there has been considerable interest in researching students’ use of corpora for error-correction and writing improvement (Gaskell & Cobb, 2004; Gilmore, 2009; Mull & Conrad, 2013; Tono, Satake, & Miura, 2014; Watson Todd, 2001; Yoon & Jo, 2014). This research took place in a variety of pedagogical circumstances: different corpora were used, including the British National Corpus (Tono et al.) and the COBUILD Corpus and Collocations Sampler (Gilmore); corrections were carried out in class (Mull & Conrad) and independently (Watson Todd), with a time limit (Yoon & Jo) and without (Gaskell & Cobb); varying amounts of assistance were provided including pre-cast links for concordances (Gaskell & Cobb), coded error feedback (Tono et al.) and underlining errors (Gilmore). Although most research dealt with both lexical and grammatical errors, Watson Todd focused exclusively on lexical misuse, while Mull and Conrad and Tono et al. dealt only with grammatical problems. Despite their differences, all these studies reported substantial figures for accurate correction of errors at the word and sentence level, with the majority recording a success rate of 60-70%. However, most research to date has targeted undergraduate writers at an intermediate level and their writing assignments have been relatively general and non-specialist in nature. It is less clear whether students at doctoral level, with more advanced English skills and writing highly specialised texts would also benefit from consulting a corpus to deal with their errors. Thus the aim of this paper is to determine whether doctoral students working on their theses make word and sentence level errors that can be corrected using corpora and if so, whether they are able to self-correct using a tailor-made corpus and with minimal input from the teacher.

This research focuses on an EAP course designed to teach corpus-assisted editing to doctoral students. The programme consisted of six two-hour sessions in which students were introduced to the AntConc software (Anthony, 2014) and each student built two tailor-made do-it-yourself corpora (Charles, 2012, 2017; Lee & Swales, 2006). The first was a corpus of research articles (RAs) in their own field and the second was a learner corpus of the draft chapters of their own thesis. The RA corpora ranged in size from 77,000 to 3.3 million words, with a median of about 500,000 words. In class, students were shown how to use corpus tools for editing and practised both on pre-set tasks and self-chosen queries. Outside class, students also carried out a ‘joint editing task’ with the teacher/researcher, in which they used their RA corpus to edit a single chapter of their thesis. This paper is based on data from 20 students who did this task.

Working on an e-copy, the teacher/researcher initiated the task by using the Comment function of Word to draw students’ attention to problems in their text and to ask them to use their RA corpus to deal with the issue. Problems were indicated by highlighting the relevant segment of text and stating ‘Problem for corpus search’ in the comment. No further hints were given as to the nature of the problem or the searches that would be appropriate. However, the teacher/researcher checked that the necessary information to amend the text was available in the student’s corpus before the problem was flagged for attention. Students were asked to amend their text using the Track Changes function and to add their own comment on the problem, their search processes and results before returning the chapter. Students’ amendments were checked and the chapter returned with
additional comments where necessary. This paper analyses the first three problems for corpus editing in each student’s text; it reports on the source of the problem, the success or otherwise of the student’s edit and their comment. In two cases no correction was made and not all students commented on all problems; thus the data consist of 60 problems, 58 amendments and 52 comments.

All participants had completed a minimum of one year of doctoral studies and written at least one draft chapter of their thesis. Only one, a student of linguistics, had prior experience of using a corpus. A wide range of disciplines was covered, with half the participants studying natural sciences, seven working in social sciences and three in arts/humanities. Students spoke nine different L1s, of which the most prevalent were Chinese (8), Korean (3) and German (3).

The students’ problems were analysed using the coding system developed by Chuang and Nesi (2006). This system has a hierarchical structure in which the major code refers to the language level of the error: grammatical, lexicogrammatical or lexical, while subcodes describe the linguistic category, e.g. determiner, preposition, countability of noun, or collocation. An indication of surface structure deviance is also included, consisting of five categories: omission, overinclusion, misformation, misselection and misordering. Thus a typical error tag has three parts: language level, linguistic unit and surface alteration. Example (1) indicates that a determiner-article ‘the’ (dtar the) has been omitted (–) and that this is a grammatical error (G).

(1) Tell-Aswad glazes are all of {dtar the – G} high-lead compositional type...

The results show that the majority of the 60 problems (58%) were at the grammatical level, while the figures for lexicogrammatical and lexical problems were similar at 22% and 20% respectively. Further breakdown of the figures reveals that the most frequent grammatical error was the use of determiners (14 instances, 23%), with omission of ‘the’ the single most prevalent problem, as illustrated above. The most frequently occurring lexicogrammatical problem concerned noun related preposition use (9 instances, 15%), almost all of which showed misselection of the preposition, as underlined below in example (2):

(2) There is a description from Tangier...

The lexical problems were more varied, although miscollocation occurred in seven instances, as seen below in the underlined misselection of the verb in example (3):

(3) ...salient peripheral cues... pull participants’ attention to that location.

These results are generally in line with those found by Chuang and Nesi (2006) for undergraduate writing. It is the ongoing persistence of such problems even at this level that renders corpus-assisted editing a valuable procedure for doctoral students.

In order to get a more detailed picture of student corpus use, both comments and corrections were analysed to determine first whether the students managed to identify the problem accurately and second whether they succeeded in correcting the problem. The results showed that most students did identify the problem correctly (93%), with only four instances of misidentification. Of course, the fact that the segment of text containing the error was highlighted certainly directed students’ attention towards the problem, making accurate identification more likely. In accordance with previous research, most problems (72%) were also corrected successfully and students’ comments also showed a process of
induction at work. Example (4) illustrates how a Korean student investigated and explained the use of the noun ‘politics’:

(4) Clusters: politics
44 hits for politics “is”
22 hits for politics “are”
Both are possible, but I am talking about general idea by “politics” in this sentence, so it should be ‘politics + singular form’.
Amended text underlined: ...how politics influences the choice of industrialization...

Although her search technique can be criticised, the student managed not only to deal with the problem, but also to formulate a guideline for her own future use.

Of the 17 amendments classified as unsuccessful, six were due to unsatisfactory searches and four to misidentification of problems; there were three instances of partial success, two without any correction and one instance each of an erroneous explanation, and a correction without corpus use. However, even unsuccessful attempts at editing can promote the noticing of useful language, as shown in this comment (5) by a Chinese student:

(5) I launched a search of “echo*” and got 19 hits. After sorting them with the logic of 1R, 2R and 3R, I noticed several collocations, namely 1) X is echoed by; 2) X echoes in/ has echoes in. So I would replace “echoing with” with “echoing in”.
Unsuccessfully amended text underlined: Echoing in the call for community level self-governance...

Although the student did not notice instances of the verb ‘echo’ with a direct object and thus did not correct the problem successfully, he did identify other important patterns associated with the verb. Moreover, his comment shows that he can make good use of search techniques and notice frequently recurring patterns in the data. Despite his lack of success on this occasion, then, he is well on the way to becoming competent at corpus-assisted editing.

This paper argues that even students at advanced doctoral level can benefit from the use of corpora for self-correction. Not only do students still make word and sentence level errors that they can correct using their corpus, but in the process they are able to induce rules and make discoveries that enhance their disciplinary and linguistic awareness.

References


Distant reading intimate encounters: a big data approach to online erotica
Alon Lischinsky and Kat Gupta (Oxford Brookes University, UK)

Once imprisoned in “secret museums” and hidden from the view of the general public (Kendrick, 1997), pornography has become an increasingly visible and important part of cultural life over the past 50 years. Visual and written representations of sexual activity, formally banned as obscene in most Western countries since the mid-19th century, entered the mass market in the 1960s and their circulation grew very significantly with the proliferation of special-interest magazines in the 1970s, the launch of home video systems in the 1980s, and the commercial internet in the 1990s (Hardy, 2009). Online pornography in particular has disrupted the various barriers historically erected to regulate access to erotic materials, giving rise to heated discussions about acceptable forms of sexual knowledge, sexual freedom and sexual representations (Atwood, 2010).

This newfound visibility has resulted in a dramatic increase in the amount and range of scholarly work on porn. While academic research on the subject until the 1990s tended to focus on alleged undesirable effects of porn consumption —such as undermining traditional values of monogamy and emotional attachment (Zillmann, 1986), or enticing men to sexual violence against women (Mackinnon 1989)— current work adopts a much more nuanced view of the various forms in which porn is consumed, of its psychological and social functions, and of its aesthetic and cultural significance (Wicke, 1991). But while this scholarship has led to increasing awareness of the various forms of pornographic expression, it has largely focused on the visual genres of photography and film, and exploration of the language of contemporary porn remains limited and uneven (Wicke, 1991:75).

Specific forms of erotic writing have received some critical attention when they intersect with established research traditions; for example, work from a lavender linguistics perspective has investigated the ways in which gay, lesbian and queer identities are constructed in pornographic genres (Baker, 2005; Bolton, 1995; Jacobs, 2000; Koller, 2015; Morrish & Sauntson, 2007), while scholars of a primarily historical and literary bent have explored the evolving shapes of pornographic genres (Frantz, 1989; Hunt, 1993; Moulton, 2000; Virdis, 2014). But, somewhat paradoxically, research on contemporary, heterosexual, mainstream pornography has been scarce (among the few exceptions, see Marko, 2008).

Such an omission is especially unfortunate because any attempt to interpret the language used to depict sexuality and sexual activity, and especially its relation to contentious social issues, is necessarily probabilistic and comparative (Partington et al., 2013:12). Any principled attempt to evaluate contentions such as that contemporary pornographic texts are characterised by “representational practices that demean women” (Jeffries, 2007:1) requires the comparison of a corpus of such texts with references built from analogous forms of language use (such as fiction in general, or even specific genres showing specific affinities such as romance fiction; cf. Patthey-Chavez et al., 1996). In the absence of a systematic description of overall patterns in erotic writing as a whole, any claims about the relevance of the linguistic phenomena observed to occur in specific cases or subtypes are essentially
speculative, and risk being biased by the drive towards corroborating previous hypotheses (Marchi & Taylor, 2009).

Equally important is that such descriptions should avoid the temptation to reduce genres as a homogeneous monolith. Partington et al. (2013:302) have pointed out that typical pairwise comparison techniques such as keywords are intrinsically geared towards identifying inter-corpus differences, and conversely neglect similarities between corpora. In a similar manner, these techniques are of little help in detecting or measuring internal diversity within a corpus, and as a result may foster the appearance of uniformity or homogeneity within predefined categories (Gries, 2006). In the light of the ongoing diversification of contemporary pornography, characterised by ‘diff’rent strokes for diff’rent folks’ (Mazières et al., 2014:80), reducing the key features of erotic writing to just a central tendency would obscure the variety of sexual and stylistic interests that underlie it.

In this paper, we employ two sets of techniques for dimensionality reduction to attempt a systematic description of this diversity. On one hand, we employ topic modeling tools borrowed from computational linguistics (Jaworska & Nanda, 2016) to produce data-driven, unsupervised classifications of stories collected from Literotica.com, one of the oldest and largest erotic fiction repositories online containing approximately 1.25 billion words. On the other hand, we employ the resampling and bootstrapping techniques described by Gries (2006, 2008) to attempt to measure the degree of homogeneity in the corpus and determine an appropriate level of granularity for classification. Using co-occurrence measures, we show that the diversity of content in the stories is not easily captured by pre-set ontologies such as the genre system employed by the repository, but rather reflects a complex mixture of semantic and stylistic traits that require detailed linguistic analysis.

References


This research will:

- Extend a corpus-based language learning platform to Chinese.
- Enable an exploratory/discovery-based learning approach for Chinese learning
- Conduct grammatical analysis of Chinese for natural language processing

Mandarin Chinese has the largest number of native speakers of all languages. Many people are starting to learn Chinese, and more resources enabling its learning are needed. Already, there is a wide range of apps and web learning platforms, targeting in particular the mastery of Chinese characters (Shen, 2005). Other software packages focus on the acquisition of vocabulary, for example through the use of online flashcards (Edge, Searle, Chiu, Zhao, & Landay, 2011). Various online dictionaries are available, some offering advanced features.

What we have not seen so far is applications which make use of or are based on authentic text data, even in online dictionaries. As was once more often the case with TESOL materials, dictionaries give example sentences which are either invented, or drawn from literary sources that do not reflect modern usage. Corpora are not widely used by Chinese lexicographers and materials developers (Li & Smith, 2015).

One particularly powerful corpus-based tool, currently available for English and Russian learning, is SkELL (Sketch Engine for Language Learning; Kilgarriff, Marcowitz, Smith, & Thomas, 2015). SkELL is a user-friendly system which allows learners to explore the behaviour of words in context, presenting a variety of example sentences, and showing how words participate in collocations and interact grammatically with other words. We will extend SkELL to the Chinese language.

With SkELL, language learners are able to consult three types of word usage display, based on large corpora. They are Example Sentences, which uses the GDEX algorithm (Kilgarriff, Husák, McAdam, Rundell, & Rychlík, 2008) to find the best dictionary examples from a corpus; Wordsketch, which gives a single-screen summary of a word’s usage, showing which other words it typically collocates with, and in what grammatical relations; finally there is Similar Words, which is a distributional thesaurus. SkELL is based on the architecture of Sketch Engine, a general purpose corpus analysis tool which is not specifically configured for language learning, but which does offer access to several large Chinese corpora. Most of these, in common with corpora of other languages available on Sketch Engine, have been segmented (divided into words) and POS-tagged. In order to provide the Wordsketch and Similar Words features, a sentence parsing function is also required, and this takes the form of a set of language-specific grammatical relations rules which are
executed on the fly (when the user requests the feature). The rules are specified as regular expressions over parts of speech.

For our implementation of SkELL, we will evaluate a range of segmenting and POS tagging options, including the tools of Academia Sinica (Ma & Chen, 2005; Tseng & Chen, 2002), City University Hong Kong (Sun, Shen, & Tsou, 1998) and Stanford University (Toutanova, Klein, Manning, & Singer, 2003; Tseng, Chang, Andrew, Jurafsky, & Manning, 2005). We will also optimize the grammatical relations rules, to take better account of the relatively free word order of Chinese (including for example topicalization and object-fronting) than does the ruleset currently implemented on Sketch Engine.

The poster will set out plans for the SkELL implementation, and will incorporate a small-scale evaluation of the currently available Chinese Wordsketch function, exploring ways in which the grammatical relation ruleset might be optimized.

References


Imperfect language learning vs. active sound change: The shift [i]>[e] in the verbal pattern hif’il in Modern Hebrew

Einat Gonen (Department of Hebrew Language and Semitic Linguistics, Tel Aviv University, Israel) and Yael Reshef (Department of Hebrew, The Hebrew University of Jerusalem, Israel)

Modern Hebrew provides an idiosyncratic case for the study of processes of linguistic change due to discontinuity of its use as a spoken language. Whereas in other languages differences between various historical phases necessarily reflect the outcome of change processes, in the case of Hebrew such differences may have two different sources: (1) imperfect implementation of the classical rules by the original L2 speakers of Modern Hebrew at the initial stages of speech revival; (2) processes of linguistic change which operated in the speech community in further generations. However, delineating the boundaries between these two different categories is not always possible, as the speech habits of the first generations of speakers are unknown since the systematic study of spoken Hebrew is relatively recent, dating back to late 20th century.

The paper offers a new research direction based on the recent discovery of two collections of recordings of spontaneous Hebrew speech made in the 1960s, which open up new possibilities for the diachronic study of spoken Modern Hebrew. One collection consists of informal interviews with first generation speakers of Hebrew (born in the first quarter of the 20th century). The other collection consists of recordings made within one family over 50 years (1963-2013), documenting daily situations in the family life. As all hitherto known historical recordings of Modern Hebrew reflected formal language use, the availability of these new collections offers for the first time the possibility to explore processes of change in spoken Modern Hebrew based on solid linguistic data.

The paper focuses on one conspicuous sound change in contemporary Modern Hebrew, namely the transition from [i] to [e] in the prefix of the verbal pattern hiCgiC, called in Hebrew hif’il. Recent studies (e.g. Bolozky 2007) pointed at the growing distribution of the variation hef’il at the expense of hif’il beyond the circumstances prescribed by the rules of classical Hebrew grammar, leading to a morphological change in the prefix used in this verbal pattern. The linguistic data extracted from the 1960s recordings clearly indicate that the variability found in contemporary language between hif’il and hef’il has two distinct sources:

(1) An initial state of variability between [i] and [e] in forms derived from weak root verbs (initial-Nun and middle-Waw/Yod roots, e.g. higi’a-hegi’a ‘arrived’) due to imperfect language learning in the initial phases of the formation of Modern Hebrew.

(2) A recent change from [i] to [e] in hif’il forms derived from regular triconsonantal roots (e.g. hitxi-l-hetxi-l ‘started’). In this category, the 1960s
recordings attest to a stable realization of the rules of traditional Hebrew grammar in all age groups, and the occurrence of deviating forms is marginal.

The following table provides the data extracted from the 1960s recordings for two generations of speakers: (1) old adults (born in the first quarter of the 20th century, in their 40s-60s at the time of recording); (2) young adults (born around the 1940s, in their 20s at the time of recording):

<table>
<thead>
<tr>
<th>Age group</th>
<th>Grammatical form: regular roots (hiCCiC)</th>
<th>Grammatical forms: initial guttural (heCCiC)</th>
<th>Non-grammatical form: regular roots (heCCiC)</th>
<th>Non-grammatical form: initial guttural (hiCCiC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Old adults</td>
<td>299 (92.57%)</td>
<td>59 (95.16%)</td>
<td>24 (7.43%)</td>
<td>3 (4.84%)</td>
</tr>
<tr>
<td>Young adults</td>
<td>96 (83.47%)</td>
<td>12 (92.31%)</td>
<td>19 (16.52%)</td>
<td>1 (7.69%)</td>
</tr>
</tbody>
</table>

Based on Ohala’s theory of sound change (Ohala 1981, 1989, 1993, 2012), the measure of synchronic variation documented in the 1960s recordings will be analyzed as a precursor of the sound change that developed in the language at a later stage.

References


Extracting English Verb-Noun Miscollcations from NNS Academic Writing based on Corpus Comparison
Hao-Jan Howard Chen and Fiona Yu-chen Cheng (National Taiwan Normal University, Taiwan)

There have been more and more studies showing that collocation played a significant role in second language comprehension and production (Ellis, 1996; Lewis, 2000; Lien, 2003). However, collocations were often found to be problematic for many second language learners. Previous studies have consistently revealed that EFL learners had insufficient knowledge of English collocations (Bahns & Eldaw, 1993; Gitsaki, 1997; Liu, 1999; Chen, 2008). Laufer and Waldman (2011) found that L2 writers produced far fewer verb-noun collocations than native speakers. Moreover, the number of verb-noun collocations errors, particularly interlingual ones, continued to persist even at advanced levels of proficiency.

While verb-noun collocations “form the communicative core of utterances where the most important information is placed” (Altenberg, 1993, p. 227), researchers indicated that the most frequently occurred miscollcations were verb-noun miscollcations, and they were particularly difficult for second language learners (Liu, 2002; Nesselhauf, 2005).

Empirical studies have offered clear evidences that EFL learners lacked verb-noun collocational knowledge and made suggestions for explicit instructions on collocations. According to Woodlard (2000) and Lewis (2000), helping students to observe and notice their own miscollcations would enhance students’ awareness of acceptable collocations. Therefore, investigating learners’ miscollcations can help teachers and researchers to better understand learners’ general patterns of collocational errors, thus shedding lights on what collocations to teach and how to help language learners.

In most existing studies on miscollcations, researchers in these studies collected L2 learners’ compositions, assignments, or examination essays, and manually marked these miscollcations. There are several studies targeting on L2 learners’ miscollcations through the use of learner corpora (Nesselhauf, 2003, 2005; Shih, 2000; Liu, 2002; Chang & Yang, 2009). One of the most comprehensive studies is the verb-noun collocation study carried out by Nesselhauf (2005). She investigated the use of verb-noun collocations produced by advanced German learners of English based on International Corpus of Learner English (ICLE). Nesselhauf manually extracted and analyzed the verb-noun combinations in the 318 essays selected from the sub-corpus (GeCLEE) which contained around 150,000 words. The study was able to retrieve 507 V-N miscollcations produced by German EFL learners. It is evident that Nesselhauf spent a great amount of time in searching through the learner corpus and provided useful information about learners’ miscollcations. It is, however, very labor-intensive and time-consuming to examine second/foreign language learners’ miscollcations through manual error-tagging. However, several much larger learner corpora are now available. These include corpora such as ICLE (International Corpus of Learner English), International Corpus Network of Asian Learners of English (ICNALE), The TECCL corpus (Ten-thousand English Compositions of Chinese Learners) and TOEFL11: A Corpus of Non-Native English. The manual error tagging and error analysis methods might not be able to process these larger learner corpora. Second language researchers need to adopt some new approach to analyze these learner corpora and to uncover various miscollcations more efficiently.

In this paper, we will introduce a useful corpus research tool called Sketch–Diff, a tool which was provided in Sketch Engine (SKE) developed by Adam Kilgarriff and his associates. This powerful tool can help language researchers to compare various
collocations used in a native corpus and a learner corpus. In this study, we first uploaded the following two corpora onto SKE: one native reference corpus and one NNS academic writing corpus. The native corpus includes the academic sub-corpora taken from BNC (British National Corpus) and COCA (Corpus of Contemporary American English) academic corpus (about 100-million words). The NNS academic writing corpus is a large collection of 494 Taiwanese masters’ theses in TESOL from 10 different graduate programs in Taiwan (10-million words). After these two corpora were uploaded into SKE, the SKE can automatically tag and analyze these corpora. After the pre-processing stage, the Sketch-Diff tool can then be used to compare all the collocations used by native speakers and non-native speakers. When the researcher input any target word, he/she can then find all collocates of the target word used in the uploaded native corpus and non-native corpus. Various collocation patterns can be extracted and the significant differences between the usage of native speakers and non-native speakers can be revealed. The system also used different colors to show the users about the significant differences between native and non-native usage.

With the help of SKE tools, 171 types (1171 tokens) of miscollocations were discovered in this study. All miscollocations were categorized into four main major types: misused verb, misused noun, misused or missing preposition, and nonexistent combination. 142 types (946 tokens) of miscollocations were categorized as misused verb; 16 types (124 token) were misused noun; 6 types (36 tokens) were misused or missing preposition and 7 types (75 tokens) of miscollocations were non-existent. Moreover, many of these collocation errors were attributed to L1 influences. 95 types (608 tokens) of miscollocations were possibly influenced by learners’ native language. It seems clear that first language plays a crucial role in second language learners’ production of collocations.

Based on our empirical tests, the Sketch-Diff can be an effective tool in comparing collocations used in English native and non-native corpora. With this new corpus comparison tool, researchers do not need to painstakingly examine each and every line of students’ essays and manually mark each and every collocation error. This new corpus processing tool provides a more convenient and thorough way of uncovering the possible differences between native and non-native’s collocational competence. At the current stage, the corpus tool is surely not fully automatic in uncovering all the collocation errors made by L2 learners. Researchers still need to further verify the differences recommended by SKE. However, the tool has already made some contributions to studies on L2 acquisition of collocations.

References


Visualizing Language Change in a Corpus of Contemporary German
Peter Fankhauser and Marc Kupietz (Institut für Deutsche Sprache, Germany)

We introduce an approach to visualize language change in a large corpus of contemporary German newspapers, spanning the years 2000 to 2015, drawn from DeReKo (Kupietz et al. 2010). The visualization combines two factors involved in language change: Semantics and frequency change.

Semantics of words is visualized by positioning them in two dimensions such that words with similar co-occurrence contexts are positioned closely together. This is accomplished in two steps: First, word embeddings are computed with the structured skip-gram approach described in (Ling et al. 2015), which takes into account word order. To calculate individual word embeddings for each year, we follow the approach of Dubossarsky et al. (2015) and Kim et al. (2014): The embeddings for the first year are initialized based on newspapers from 1950 to 1999 the embeddings for each subsequent year with the embeddings of the previous year. With this approach, the embeddings are comparable across years. Second, the 200 dimensions resulting from the first step are further reduced to two dimensions using t-Distributed Stochastic Neighbor Embedding (Van der Maaten & Hinton 2008).

Frequency change is represented by color, ranging from violet for words with decreasing frequency to red for words with increasing frequency. To this end we calculate the slope of the generalized linear fit of a logistic transform on the relative frequencies in each year (Zuraw 2003) and map it to the color range. The goodness of fit for linear fits is 44.4% and for 2nd order fits 59.2%, thus, the colors characterize frequency change fairly well.

Figure 1 gives an overview on the visualization. To the left, a bubble chart represents the color encoded semantic space of words, with the size of bubbles...
proportional to the square root of the relative frequency in the chosen year (here: 2015). To the right frequency change of individual words is represented by simple line charts showing the fitted 2nd order polynomials of the logit transformed relative frequencies. The line chart also doubles as a selector for individual years.

Figure 2. Two regions zoomed in

Figure 2 shows two zoomed in regions, female first names to the left, and brand names of tech companies to the right. In both cases we can see that frequency change is typically correlated with semantic similarity, i.e., words close to each other typically have a similar frequency slope.

Other uses of the visualization include detection of semantic change and replacement of obsolete words. The visualization has been implemented using the D3 framework (Bostock et al. 2011), postprocessing of word embeddings and fitting of the generalized linear models in R. The visualization is publically available at: http://corpora.ids-mannheim.de/diaviz/dereko.html

References


Computer Assisted Legal Linguistics (CAL²): An interdisciplinary approach

Yinchun Bai (University of Freiburg, Germany), Isabelle Gauer (University of Freiburg, Germany), Hanjo Hamann (MPI Bonn, Germany) and Friedemann Vogel (University of Freiburg, Germany)

Introduction

In law as well as in linguistics researchers work with texts. In recent years both disciplines are turning away from the introspective approach to a more empirical, evidence-based practice. Digital sources and programmes for statistical analyses of mass data have become increasingly available and offer new opportunities for linguistic research. In law and legal linguistics a growing number of researchers and practitioners are also acknowledging the merits of working with large text collections, so they promote steps towards an evidence-based jurisprudence (Goźdź-Roszkowski, 2011; Mouritsen, 2011; Hamann, 2014; Fagan, 2016; Solan, 2016; Vogel, 2017). “Computer Assisted Legal Linguistics” (CAL²) (Vogel, Hamann, and Gauer, in press) tries to support the process of legal problem-solving by using corpus linguistic methods. That means, computer-supported analysis of carefully pre-processed corpora of legal texts are used to analyse legal semantics, language and sociosemiotics in different working contexts (judiciary, legislature, legal academia, etc.). It does not seek to replace hermeneutic procedures with algorithms but to complement them. Empirical data and computer algorithms can only support legal work, but decision making will always be cognitive processes of contextualization.

The CAL² corpus of European law

The CAL²-group seeks to bring together members from the fields of law, linguistics and computer science and to build an infrastructure for legal linguistic analysis. To do this, the first step was to create a corpus that is suitable and balanced enough to address questions on the language and law interface. That means, it must contain texts that exhibit the special properties of the legal genre. At the moment, we have included German court decisions, statutes and articles from legal journals and English court decisions as the starting point of building a corpus of European Law. To be more specific, the CAL² corpus of European law now contains: 6,300 German federal statutes (~15 M words), 370,000 court decisions in German (~800 M words), 20,000 court decisions in English (~90 M words) and 43,000 German academic research papers (~200 M words), which equal over 1 billion words in total (see Figure 1 below for the temporal distribution of the data).
The collected texts were converted to TEI P5 (www.tei-c.org/Guidelines/) compliant xml which serves as a de facto standard for text structural annotation (Stührenberg, 2012). A pipeline of xsl transformations tailored to the different types of input formats was created to construct the corpus. Additionally, we enriched the data with part-of-speech annotation using TreeTagger (Schmid, 1995). During the corpus building, error control and duplicate removal were integrated in the process. It was important to get the data as clean as possible to be able to conduct quantitative as well as qualitative analysis. To assist the cleansing work and to get a better overview of the corpus content, the metadata of the legal texts were also collected and documented in a related database.
Pending steps

The next step is the development of a platform for specialised statistical processing that supports legal linguistic research and further also legislation and decision making (for example statutory interpretation of ambiguous expressions). The platform will help to gain insight to questions like how legal terms were used in the past and today, how often and by whom. This can assist the research on diachronic changes of linguistic expressions in law and the development of dogmatic schools.

The implementation of the platform will comprise three parts: First, frequency lists will be created to derive the 200,000 most frequent lemmas (nouns/verbs/adjectives) of the corpus. For all of them we will calculate context profiles for statistical multi-level-context-analysis that contain co-occurrences and usage patterns according to the metadata (for example when the expression was used most frequently). Secondly, we attempt to measure the rigidity of expressions to calculate how fixed or varied the expressions are in different contexts. Lastly, we examine semantic similarity in legal texts by clustering similar context profiles and visualizing the results as self-organizing maps.

Case study

To test the CAL² corpus in action and to explore its applicational possibilities and limitations, we conducted a case study using parts of the corpus data. The case study is a comparative study of the linguistic formulation of the “employee” concept in UK and EU court decisions. The choice of the term “employee” as our study object was motivated by several considerations. First, the employment law is one of the most dynamic legal domains. Since it does not regulate the behaviour of the legal subjects directly, but confines and develops individual rules which in turn regulate the relevant affairs, the working techniques involved are complex and hardly have legal benchmarks to compare with (Vogel, Pötters, and Christensen, 2015). The development of terminology in this legal domain is therefore especially complicated, which made the study of an employment law-related notion highly interesting. Second, the term “employee” is an especially important concept in the framework of the employment law, because the employment law is by nature the employee protection law (Hueck and Nipperdey, 1957) which mediates the relationship between employees, employers, trade unions and the government.

Data

Our research material of this study comes from the English part of the CAL² corpus. It consists of judgements made by the UK Employment Appeal Tribunal [UKEAT] and the employment law-related judgements from the Court of Justice of the European Union [EUECJ]. Using texts of judgement to study the use of a certain term underlies the representativeness of our findings. On one hand, the terms used in judgements are linguistically produced after conceptualization and interpretation by judges, who represent the understanding and use of a term both as a regular native speaker and as a legal professional – they are aware of the definition and the legal force of a term under the legal context but are still confined by the limitations of individual intuition about the frequency and the pragmatic implications of the usage. This helps to reflect the conception of “employee” at the dynamic user level. On the other hand, court
decisions are a specialized type of text production, which are highly contextualized and constrained by the institution of law. The formulation of language involving the use of “employee” is thus subject to specially institutionalized format as well as semantics according to the corresponding legal conventions. They therefore help to reveal the conception of “employee” at the underlying institutional level.

Research questions

This study addresses the following research questions: (1) What are the central “employee” concepts and “employee”-related concepts within the labour law framework in the UK case law system? (2) How are the “employee” concepts addressed in UK court decisions in terms of frequent usage patterns? (3) How similarly and differently are the “employee” concepts linguistically formulated and presented in EU court decisions?

Results

In terms of the frequency distribution of various “employee” concepts and the usage patterns of the term, we found both commonalities and differences in UK and EU court decisions. This is achieved through analysing the compounding and phrasal structures, the genre-specific co-occurrence partners, and the predication patterns of the term “employee”. In both UK and EU court decisions the majority of compound nouns of “employee” concern the status of a person in an employment relationship, i.e. if a person can be classified as an employee at all (e.g. ex-employee), although the EU court decisions exhibit a much lower rate of using compound nouns in general. For noun phrases containing “employee”, we identified four phrasal structures, including modifier-noun phrases, noun-noun phrases, possessive-noun phrases, and of-possessive phrases. Generally speaking, the modifier-noun phrases either identify the types of an employee in relation to the employment relationship or describe aspects of the employee proper, such as “full-time employee” and “female employee”; the noun-noun phrases, possessive-noun phrases, and of-possessive phrases, on the other hand, mostly address the expansive aspects of an employee, including e.g. people related to an employee (“employee representative”), and external components and events involved in an employment relationship (“employee’s contract”, “dismissal of employee”). From the semantic perspective, the “employee” concept in both UK and EU legal cultures has a core semantic composition of “temporal specification of an employment relationship”, “discrimination between social groups (gender, age, etc.)”, and “rights and entitlement”, as reflected by the frequent co-occurring partners in phrasal structures, expanded contexts [+/-15], and predication structures. But the UK and EU court decisions do show different emphases on a few specific aspects of the employee and the employment relationship. For example, the UK court decisions contain more predications describing an employee’s conducts and conditions in an employment relationship, while the EU court decisions feature a slightly bigger percentage of notions that concern the support of an employee.

Conclusion

To summarise, the CAL² research group engages itself in corpus-assisted legal linguistics and strives to contribute to this young and vibrant working field by compiling
a comprehensive and balanced specialised corpus of legal texts and developing accompanying analytical tools. Our approach aims at making the possible interpretations more visible and thus making the law more transparent by algorithmically searching for and analysing recurrent speech patterns in large linguistic corpora of legal texts. With a case study analysing the conception and usage patterns of the legal term “employee”, we showed one of the practical uses of the CAL² corpus and experimented on the possibility of identifying conceptual sediments that exist as constants in the European way of legal thinking.

References

Prosody-based clustering of spoken corpus search results
Piotr Pęzik (University of Lodz, Poland)

The role of prosodic features of speech in the construction of meaning and discourse structure is an active area of linguistic research (Wichmann et al. 2009). The work described in this paper addresses the need for corpus-based tools for studying prosodic phenomena in conversational and read speech. More specifically, the paper presents an experimental prosody-based clustering functionality of the Spokes online search engine available for spoken Polish and English data (see: spokes.clarin-pl.eu, pelcra.clarin-pl.eu/SpokesBNC/ and pelcra.clarin-pl.eu/spokes2-web). The basic goal of this technique is to facilitate the analysis of concordance results according to selected features of the speech signal associated with word spans matching user queries. For example, given a simple query for the word hello, it is possible to automatically label pitch patterns detected for this surface form and thus distinguish between rising vs. falling pitch contours or relatively high- vs. relatively low-pitched instances of hello in the corpus. The assumption behind this approach is that even a simple sorting of search results by such labels may lead to the discovery of prosody-marked functions of words and multiword expressions.

Furthermore, normalized annotations of prosody can also be used as more complex feature vectors in order to cluster concordance result sets in a multidimensional space. The characteristics of speech considered in this experimental module of Spokes are a) duration of the matching search spans, b) pitch contours and c) voice intensity measurements. The normalization of pitch and intensity contours is performed using the MOMEL technique (Hirst & Espesser 1993) for simplifying contours and the INSINT scheme for encoding intonation patterns. The applicability of this new data exploration method to corpus-based studies of spoken language is evaluated on two use cases a) detecting potential prosodic distinctions between instances of ditropic expressions (cf. Van Lancker et al. 1981) and b) identifying functionally different but morphologically identical discourse and pragmatic markers.

References


An analysis of stance and voice in Applied Linguistics research articles across Mainland Chinese and British cultures

Xiaoyu Xu and Hilary Nesi (Coventry University, UK)

Scholars from Mainland China are increasingly publishing in the medium of English in order to gain visibility and credibility worldwide. However, while Chinese scholars in the hard sciences now feature extensively in international databases such as the Science Citation Index, the visibility of Chinese scholars in the Social Sciences Citation Index is strikingly low. Due to the holistic, interpretative, reiterative nature of knowledge in the soft sciences, writers have to work harder to establish personal credibility through their claim-making negotiations, sharing sympathetic understanding and promoting tolerance in their readers (Becher, 1994; Becher & Trowler, 2001; Hyland, 2000). The Chinese language has developed in a strict political environment within the particular cultures of Confucianism and collectivism where high importance is given to context and greater energy is devoted to saving ‘face’. It is probably for this reason that the tradition of argument in China is considered to be more implicit, more modest, more positive, less engaging and less evaluative than in the Anglo-American argumentative traditions (Peng & Nisbett, 1999; Lustig & Koester, 2008; Hofstede & Minkov, 2010; Hall, 1976). However factors other than national culture such as disciplinary culture, genre, topic, audience size and the writers’ own language proficiency and academic experience may also impact on argumentative style in research writing. These considerations have been neglected in most of the prior research into the typically Chinese features of academic writing.

This paper uses corpus methods alongside discourse analysis to test whether and how all these factors might affect research writing in the soft sciences. To this end, 15 Applied Linguistics research articles written in English by Mainland Chinese scholars were selected for a Chinese corpus, and 15 research articles by British scholars matching the variables in the Chinese corpus were selected for a reference corpus. Only the Introduction and Conclusion sections were included as they are the most argumentative parts of research articles. Stance and voice, two crucial aspects of argument, were chosen to be examined using Appraisal Theory (Martin & White, 2005) which maps the expressions of feelings, the interaction between reader and writer, and the up-scaling and down-scaling of evaluations. Appraisal markers were differentiated according to their level of relevance to the writers’ argumentative intention, using the theory of context (Xu & Nesi, 2017). This helped to focus on Appraisal markers relevant to argument. Moves were also taken into account, drawing on Genre Analysis (Swales, 1990, 2004) to understand the purposes of particular co-articulations of Appraisal markers. The two corpora were marked up manually using the UAM CorpusTool (O’Donnell, 2008) and compared statistically using the t-test provided by UAM CorpusTool.

The findings indicate that generally the Chinese authors tried to maintain writer-reader relationships by avoiding explicit attitudinal evaluation of the work of others, while the British authors tried to maintain writer-reader relationships by toning down or only evoking stance. The Chinese authors argued for their own positions by reinforcing their explicit attitudes, adding multiple references, sharpening the completion of tasks and construing claims as unquestioned, whereas the British authors argued for their own positions by explicitly evaluating people and phenomena (see Table 1).
Table 1. Chinese and British authors’ characteristic preferences for Appraisal categories
(+ = weak significance 90%; ++ = medium significance 95%; +++ = high significance 98%)

<table>
<thead>
<tr>
<th>Chinese</th>
<th>Argument for own positions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maintenance of writer-reader relationship</td>
<td>Up-grading of Inscribed Attitude +++</td>
</tr>
<tr>
<td></td>
<td>Multiple references +++</td>
</tr>
<tr>
<td></td>
<td>Sharpening +++</td>
</tr>
<tr>
<td></td>
<td>Monoglossic claims +++</td>
</tr>
<tr>
<td></td>
<td>Expression of Attitude ++</td>
</tr>
<tr>
<td>Heteroglossic claims +</td>
<td>Inclination +++</td>
</tr>
<tr>
<td>Evoked Graduation ++</td>
<td>Reaction ++</td>
</tr>
<tr>
<td></td>
<td>Social valuation +++</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For example, while trying to avoid expressions of explicit Attitude to downtone stance and voice, the Chinese writers in this corpus consistently preferred to make Monoglossic claims (e.g., *Compared with the previous quantitative methods in typology, our method has these advantages and novelties*). These claims are presented as “taken for granted” and are “intersubjectively neutral, objective or even ‘factual’” (Martin & White, 2005: 99). The Chinese writers were more inclined to establish territory in their introductions through Multiple references (e.g., *The same controversy exits in the EFL field among the relatively sparse studies on PA training and its long term effect assessment (Lundberg et al., 1988; Bradley & Bryant, 1985; Lie, 1991; Kozminsky & Kozminsky’s, 1995; Castles & Coltheart, 2004)*) and through Sharpening the reality process (e.g., *research has shown that..., studies have found...*), referring to the prior research to evoke the importance of the topic and an authoritative stance. They also strengthened their stance by up-grading Inscribed Attitude (e.g., *early childhood literacy is the best investment*) to emphasize potential improvements that could be made to the real world. In the concluding section they contextualised their studies by Sharpening the Focus (e.g., *...as demonstrated in this study..., this study has shown*), signalling that their aims had been realised. When they consolidated their results, they up-graded Inscribed Attitude (e.g., *a key strategy in this would be for tutors to...*) to emphasize the positive implications of their research for the real world.

In contrast, the British consistently preferred Heteroglossic claims to generally downtone their stance and voice, making reference to other voices by a variety of means such as projection (e.g. *these findings suggest that...*), modality (e.g. *the criticism of PPP seems well-founded*) or concession (e.g. *Although generally unstated in descriptions of methodology, this belief ...*). However the British writers established territory by using more explicit Attitude (e.g., *the latter two points are particularly concerning*) to evaluate important problems mentioned in the literature, evaluate the advantages of prior approaches, and reluctantly agree with opponents’ opinions. When they moved to establishing a niche, they preferred Inclination (e.g., *I am interested, we hope, we are curious*) to directly reveal authorial presence and their desire to explore the real or research world context. Stance was more balanced in the British writing, with greater use of Evoked Graduation to Soften the reality of the phase (e.g., *suggest, sought to, belief*) so as to evoke the negative value that the prior research was incomplete, or to
evoke a dialogic space for disagreement with the writers’ suggestions or opinions. They also used more Evoked Graduation to predict results (e.g., the optional feedback group was expected to utilize feedback to a lesser extent and perhaps view it more casually), or to increase negative evaluations (e.g., their perception of feedback was unclear) or the importance of the research focus (e.g., contribute, enhancing). In their conclusions, Evoked Graduation, particularly upgraded Non-specific numeration (e.g., the current study thus yielded several kernels of information), was used to consolidate results in order to evoke a positive value of productivity. Evoked Graduation (e.g., such, more specifically, specific) was also used by the British writers to evoke a contractive stance and suggest specific ways to improve the real world.

The findings suggest that both the Chinese and the British authors are aware of the need to argue for their own opinions and maintain good relationships with their readers, but choose contrasting ways to realize these same purposes. These findings provide evidence to challenge the hypothesis that Chinese writers are more implicit, more modest and more positive, but they also support the view that Chinese writers are less evaluative and less engaging. Therefore, this paper argues that although culture might affect rhetorical choice, we need to explore beyond the simple cultural stereotypes (e.g., that Chinese writers devote greater energy to saving ‘face’).

Because the statistically significant differences in stance and voice strategies revealed in this paper indicate differences between Chinese and British scholars’ argumentative styles, they suggest the need for a new way of perceiving Chinese ethnolinguistic impact on research writing, and might also inform the teaching of academic writing in the social sciences.

References

Written Proficiency in English for Specific Purposes: a Corpus Study of Abstracts in Health Sciences

Ana Luiza Freitas (Federal University of Health Sciences of Porto Alegre, Brazil) and Maria José Bocorny Finatto (Federal University of Rio Grande do Sul, Brazil)

This research explores the development of written proficiency in English (Biber, 2006) for the production of abstracts in health sciences. It also aims at exposing the structural organisation and the sense-making derived from the lexical fragments which make up academic texts. For such purpose, sequences of words perform a fundamental role, both in conveying discursive identity to academic discourse, and in weaving fluent written productions (Biber & Conrad, 2009). This way the investigation aims at contributing to the advances in the studies of English for Specific Purposes (Hyland, 2008) by providing support to the creation of instructional materials and to teacher development, as well. Based on Corpus Linguistics (McEnery & Wilson, 2001), Linguistics for Specialized Languages (Hoffmann, 1987) and English for Academic Purposes (Swales, 1990), the research put together, described and analysed a corpus of 180,170 words, comprised by abstracts in Medicine, Nutrition and Pharmacy. The texts originate from academic productions from four different Brazilian universities, contrasted with abstracts from three major Anglophone journals in the same fields of knowledge. The analytical study units are lexical bundles (Biber, Conrad & Cortes, 2004) for whose extraction and identification through AntConc (Anthony, 2014) an extent criterion of 4 graphic words and a frequency and minimum distribution of 5 occurrences, in at least 5 different texts in each of the two parts of the corpus were established. 96 lexical bundles were extracted from the Anglophone portion of the corpus, which adds up to 90,098 words, whilst 88 recurrent word sequences were obtained from the Brazilian part of corpus, which amounts to 90,072 words. Regarding the metrics of lexical frequency and variability, the two data segments uncovered subtle distinctions which reveal peculiarities in the ways of building a scientific narrative by Brazilian authors when contrasted with internationally-published authors. As such, a larger repetition of word associations and a higher use of lexical bundles to express purpose and to highlight the achievement of the academic endeavour were noticed in the Brazilian corpus. The Anglophone-originated corpus, on the other hand, features more diverse recurrent strings of words, a concise prose and the use of extended collocations to highlight the scientific enterprise in itself. A detailed analysis of those distinctions leads to a realisation of the potential of scientific texts to provide evidences not only about the investigations in their experimental character perse, but also about the scientific narrative in itself, as well as about how scientists can be perceived in the international community and, as such, about how much the use of language enables them to belong or not to a given community. In the final part of the investigation, suggestions for the
application of the findings in the form of teaching tasks are provided. A virtual learning environment under construction to help users of English as an additional language produce abstracts in health sciences is also presented.

References


One of the challenges facing corpus creators and users is the fact that so many corpora quickly become "stale". They may do a great job of representing the language from 10--20 years ago, but there is nothing from the last year or two -- or especially the last month and definitely not from yesterday (see Davies 2011).

In May 2016 we released the NOW corpus ("Newspapers on the Web"), which is meant to help fill this gap (http://corpus.byu.edu/now/). The corpus is composed of more than 4.3 billion words of data from newspapers from twenty different countries, from January 2010 to the current time. (Because the corpus is continually growing, it will be more than 4.7 billion words in size by the time the CL 2017 conference is held in July 2017.)

Most importantly, however, automated scripts run every day to add texts to the corpus. Each day these scripts 1) get 10,000--15,000 URLs from Google News 2) download the web pages with HTTrack 3) clean them up with JusText to remove boilerplate material 4) tag and lemmatize the texts with CLAWS 7, 5) remove duplicates (based on n--grams), and then 6) integrate the texts into the existing relational database architecture.

In this way, there are 4--5 million words from approximately 9,000--10,000 new texts every day. For example, this abstract is being written the morning of April 20, 2017 and the corpus already contains 5.5 million words (in 11,700 articles) from yesterday -- April 19, 2017. This daily expansion of the corpus translates into about 130 million words each month, and about 1.5 billion words of data each year.

With the NOW corpus, users can see what is happening in the language this week ---- not just 10 or 20 years ago. For example, they can find the most recent 100 hits for any word or phrase, and in many cases they will have hits from yesterday. Such data is typically much more relevant and interesting to language learners than data from the early 1990s, before they were even born.

The following are just a handful of words that probably would only be in a corpus that includes very recent texts, and all of which have many tokens in NOW: (general words) Brexit (n) 2015, manspreading (n) 2015, makerspace (n) 2015, gig economy (n) 2015, dadbod (n) 2015, momager (n) 2015, swatting (n) 2015, walkscore (n) 2015, trigger warning (n) 2014, mommy porn (n) 2014, normcore (n) 2014, listicle (n) 2014, sufferfest (n) 2014, and catfishing (n) 2013;; (computers / technology / science) Uberization (n) 2015, selfie stick (n) 2015, fracklog (n) 2015, digital detox (n) 2015, droneport (n) 2015, data lake (n) 2015, smartwatch (n) 2014, and airpocalypse (n) 2013. Hundreds of examples of similar recent words can be found at the NOW website.

Of course the corpus allows users to do more than just search for individual words and phrases. They can find families of new words, such as words containing *fest (more than 2,600 different types), e.g. horrorfest, borefest, smashfest;; *sexual* (more than 1,600 types), e.g. metrosexual, demisexual, pansexual;; *phobia (more than 900 types), e.g. photophobia, dronophobia, robophobia;; smart* (more than 2,800 types), e.g. smartscooter, smartwater, smartart;; *ware
(more than 600 types), e.g. adware, crapware, Slackware;; *athon (more than 780 types), e.g. laughathon, funathon, stavathon;; *geddon (more than 230 types), e.g. Carmageddon, Stormageddon, Haimageddon. and *alypse (nearly 400 types), e.g. apocalypse, snowpocalypse, zombocalypse. Data such as this allows us to look at morphological productivity and lexical creativity in real time, since the corpus is never more than 24 hours out of date.

Because the BYU corpus architecture allows users to easily and quickly compare between different sections of the corpora, in just 2--3 seconds a researcher could find, for example, words occurring with digital NOUN that are much more common in 2015--2017 than in 2010--2012, e.g. digital nodes, digital insights, digital mall, digital factory, digital maturity, and digital chaos. A similar search for data NOUN in 2015--2017 vs 2010--2012 would yield data front, data sale, data leak, data fabric, data ingestion, or data dude and hundreds of other new phrases.

In addition to seeing the frequency in six month segments, users can also see the frequency of words and phrases by week, to see when a particular topic was discussed the most since 2010. They can also find the keywords for a given day (including yesterday), week, month, or year. For example, they could find the keywords for Apr 4 2016 (related to the Panama papers leaks), Mar 22 2016 (the Brussels bombing and Obama going to Cuba), and Nov 13 2015 (the Paris attacks).

With updates to the architecture and interface of the BYU corpus interface in May 2016, it is now possible to quickly and easily create "virtual corpora" within a given corpus (like NOW). For example, in just 5--10 seconds users could create virtual corpora based on the following words: investment*, rebound (i.e. basketball), or electron. They can also create virtual corpora based on date and source (e.g. New York Times in Sep-Dec 2016).

For example, in just 10--20 seconds, a user could create a 600,000--700,000 word corpus based on texts from the UK in September 2015 dealing with refugees in Europe. The user could then search within the virtual corpus to find any word or phrase, such as asylum, or to find the collocates of a given word. They could also compare the frequency of a word or phrase in multiple virtual corpora, such as "refugee" corpora from the US, UK, Canada, and Australia in the same time period. Finally, in just 5--10 seconds they can generate "keyword" lists from the virtual corpora, e.g. refugee, migrant, asylum, seeker, influx, border, quota, crisis, fence, camp, arrival, boat, and shelter. Obviously, this allows researchers to quickly and easily obtain data on very contemporary topics.

In summary, a corpus like NOW allows us to move far beyond moderately-sized 10--20 year old corpora, to examine billions of words of data and see language change as it occurs.

References

Corpus tools for an Academic Writing course: a case study

Nicole Keng (University of Vaasa, Finland)

This study reports on an academic writing course in which the use of corpus tools was embedded. The tools used are AntConc (Anthony, 2014) and Sketch Engine (Kilgarriff et al, 2014). The students’ reactions to this type of learning experience are discussed to show that using corpus tools can profitably be embedded in the course design in a Finnish university context.

It is well known that students from the Nordic countries have fairly good proficiency in languages, particularly in English. At the university level, the emphasis in English teaching is increasingly on academic skills. Although many university students are fluent and proficient users of English, their level of English is not as strong on the academic register as might be expected (Henriksen and Danelund, 2015). In a longitudinal, corpus-based study of Swedish students' academic writing, Shaw (2004) found that style and vocabulary development were areas in which students benefited from additional training. The main aim of the academic writing course in Finnish universities has been to assist students in writing their thesis by providing feedback from the teachers instead of training students how to create, adapt and fine-tune their own texts, through the analysis of expert-written texts. This study aims to find out how to utilize the above-named tools in a discovery-based approach to support the students’ writing process.

I will describe the stages of this study, including teaching students how to use corpus tools. Examples of students’ writing portfolios which describe how students use corpus tools in class and in self-study and their reflections about the course will be shown. Feedback questionnaires will be employed in this study to collect quantitative results showing what functions are favoured by the students.

References


Using corpora to map language: Geographical Text Analysis of UK poverty
Laura L. Paterson (The Open University, UK) and Ian Gregory (Lancaster University, UK)

This paper demonstrates the viability of using techniques from Geographical Text Analysis (GTA) on multimillion word corpora. It emphasises the fruitfulness of including geography as a variable within corpus-based discourse analysis. The paper centres on a case study comparing media discussions of UK poverty with official deprivation statistics. Using corpora of the Guardian and the Daily Mail from 2010-2015, the analysis shows that media reports of UK poverty are London-centric and tend to systematically exclude rural areas. There is also evidence to suggest that poverty becomes particularly newsworthy – thus generating more hits in a corpus – when it occurs in places that would not be expected, such as relatively high-income areas. By comparing the geographical spread of media discussions of poverty (and 85 related query terms such as unemployment, benefits, pay, austerity) we can see how each newspaper locates discussions of poverty in geographical space. The analysis also contrasts the corpus-generated results with existing official statistics of poverty to determine whether the linguistic data and the statistical data tally.

To explain the method more fully, Geographical Text Analysis (GTA) is a way of analysing language which focuses on the identifiable geographical locations referred to in texts. It allows the researcher to visualise their corpus cartographically. Once a suitable corpus or corpora have been compiled, query terms are selected to reflect the topic under investigation. For the current study these were generated by combining the top 50 lexical collocates of *poverty* in both corpora, with the addition of further terms based on wider reading (for example, the names of particular UK welfare payments like Jobseeker’s Allowance (JSA) were included in the analysis). This led to a total of 86 queries which were run on both corpora. Concordance lines were generated, downloaded, and fed through a software programme known as a geoparser. In particular, the Lancaster University adaptation of the Edinburgh geoparser (Grover et al. 2010) was used for this research.

Geoparsing involves the automatic detection of place names occurring within a set span L/R of the concordance node (+/-10 for this study). Place names can be detected at country level down to district level for locations all over the world. Thus, although the focus here is restricted to UK place names only, GTA is a versatile method which can be used to analyse the use of locations globally. Once place names have been identified in the concordance lines, these Place Name Co-occurrences (PNCs) are tagged with the geographical coordinates corresponding to their location on the Earth’s surface. The results of the geoparsing process are then manually analysed to spot errors and address any ambiguities where a place name could refer to more than one geographical location (e.g. Boston, Lincolnshire or Boston, Massachusetts). Once any erroneous hits have been eliminated, maps of the locations occurring in the context of the query nodes can be plotted (Figure 1).
There were 70,397 hits for all query terms in the *Guardian* and 463,278 in the *Daily Mail*, which were condensed during manual analysis to 6794 and 6605 hits respectively. The large differences between the geoparser output and the final hits can be explained by three factors:

1) erroneous geoparsing hits either for places outside of the UK (Albania, Toronto, Morocco) or for fictional places (Hundred Acre Wood, Downton Abbey),
2) erroneous query hits, such as *Chavez* being picked up by the query *chav*,
3) hits which upon closer inspection of the surrounding discourse were not related to poverty (e.g. the query *expens* generated hits relating to the MPs’ expenses scandal in 2010).

It is possible to train the geoparser and, as such, the results of the manual analysis of the *Daily Mail* and *Guardian* corpora will be used to decrease the rate of Type 1) errors in future work. Type 2) errors are easily removed from the geoparser output, but can be decreased pre-geoparsing by adapting queries based on an initial analysis of concordance lines. Finally, Type 3) errors are more difficult to detect pre-geoparsing, especially given the size of the two corpora, but could potentially be decreased by restricting query terms to bi-directional collocates and/or selecting or rejecting query terms based on analysis of a thinned sample of the geoparsed output. Testing these methods of reducing errors is designated for future work.

Forms of GTA have been successfully used to provide novel interpretations of literary texts and travel guides (Donaldson et al. 2015) as well as historical documents (Murrieta-Flores et al. 2015; Porter et al. 2016a), and work is ongoing with early UK newspapers (Porter et al. 2016b). The present paper is part of a project which uses GTA in combination with CDA to analyse the discursive construction and use of poverty discourses in twenty-first century mass media texts. Our choice to focus on poverty addresses the fact that statistical information alone cannot lead to in-depth
understandings of social phenomena. The primary measurements of poverty and statistical deprivation used to set government policies are quantitative. However, poverty is not merely an economic circumstance. Jo (2013: 522) argues that the relationship between poverty and shame is constructed ‘from the dominant discourse’ and cultural norms, which are ‘collectively assembled by multiple institutions which are governed by those with power and influence’. Thus, the measurement of poverty is not the measurement of an objective phenomenon. Linguistic interpretations of the term ‘poverty’ include Kress’ (1994:28) argument that poverty is ‘a characteristic which acts as a description of a person, a classification’ (1994:29). By conceptualising poverty as a label for a particular group we can begin to see how its use in discourse could be a powerful move; to label someone as ‘poor’ or ‘in poverty’ is to imbue them with a set of (negatively-loaded) characteristics that, presumably, they cannot escape. Linguistic analysis of poverty – in the form of GTA – demonstrates that, although similar lexical resources may be used by each newspaper, their use of place names locates their discussions differently, and, furthermore, the places named in each newspaper do not necessarily correspond to the statistical data.

References


The applicability of movies in legal language teaching: Evidence from Multi-Dimensional Analysis
Pierfranca Forchini (Università Cattolica del Sacro Cuore Milano, Italy)

For decades, investigations of Courtroom Discourse have inspired the works of scholars from various disciplines. Since interest has been shown from such a variety of sources, no general consensus has emerged about the perspective which is best suited to such investigations. From a linguistic point of view, the vast majority of these works has concentrated on the language in the courtroom either in terms of interaction (Beach, 1985), dynamics (Luchjenbroers, 1991), and discourse (Pridalová, 1999), or in terms of the traits marking this type of language (Garzone, Miglioli & Salvi 1995), such as the use of particular verb tenses (Beach, 1985) or of well as a discourse marker (Innes, 2010). Linguists have also been attracted by the power of language (Gaines, 2002), the role of the presentational style (Conley, O’Barr & Lind 1979) or the role of lawyers and justice (Greenfield, 2001; Silbey, 2001) in the courtroom, and they have also been concerned with narrative and story telling, or how reality is reconstructed in the courtroom (Bennett & Feldman, 1981; Pridalová, 1999; Heffer, 2010). Furthermore, given that trials are made up of different sessions, the specific moments characterizing them, such as cross-examination or the informal talk in sidebar sessions (Gaines, 2002) have also been the object of many studies.

More recently, scholars from fields such as social theory, literary theory, film studies and linguistics (cf. Machura & Ulbrich, 2001), have shown interest in the interrelation between the law and the movies. Some scholars, for example, have considered cultural factors such as the worldwide influence of Hollywood courtroom movies (Machura & Ulbrich, 2001) or the role of courtroom drama as an expression of American popular culture (Kuzina, 2001); others have taken a more historical perspective by depicting the development of American criminal trial films (Rafter, 2001); and others again have focused on the concept of courtroom justice as a genre (Silbey, 2001). The different labels present in these works, which also highlight the difficulty of defining movie genres, well illustrate this absence of agreement: Kuzina (2001), for instance, describes ten basic groups related to movies and the law such as: courtroom whodunits, legal thrillers, historical courtroom dramas, true crime courtroom dramas, lawyer films, courtroom satires and courtroom comedies, court-martial films, social issue courtroom dramas, hybrid courtroom dramas, and jury room dramas. Rafter (2001), instead, uses the label American trial films to include criminal trial films, criminal court films, and military criminal trials. Similarly, Silbey (2001) and Machura & Ulbrich (2001) mention, respectively, the trial film genre (Silbey, 2001) and Courtroom movies or law-related movies (Machura & Ulbrich, 2001) when referring to movies and the law.

Despite the considerable enthusiasm that the scientific community has shown in the various fields linked to courtroom drama, none of the above works seem to have explored the textual dimensions of Courtroom Discourse in order to verify empirically its textuality and linguistic features. This is what the present paper sets out to examine: by applying Biber’s (1988) Multi-Dimensional Analysis to trials, the primary aim is to explore their textuality. Then, starting from the assumption that movie conversation reflects face-to-face conversation, as recent studies have demonstrated (Forchini, 2009a, 2010a, 2012a, 2013a), the second aim is to verify whether, as maintained by Machura & Ulbrich (2001, p. 118), it is really “beyond dispute that the cinematic portrayal of the American
legal system and its personnel is far removed from legal reality”.

The idea behind the study is that if no significant linguistic and textual variability is found between the two investigated domains and if it is confirmed that the linguistic similarity of movie and naturally-occurring conversation is also present at a more specialized level, then it will be plausible to assume that movies are also a rich source not only for teaching the general usage of face-to-face conversation, but also for illustrating the features of more specialized language(s), such as courtroom discourse. The reason for giving such importance to movie language derives especially from the power that movies have to evoke student interest and motivation: it has been shown (Forchini, 2012b, 2013b) that students not only appreciate using them in learning environments, but also increase their linguistic competence by studying with movies. Interestingly, the impact that movies have on students was already pointed out more than 90 years ago: Cunningham’s (1923, p. 489) results proved “that the interest created by the prospect of the moving picture caused the class to work hard during the entire month”. Similarly, during the 40s, Mallery (1948, p. 149) reported that students’ found studying novels “much more interesting” combined with the use of movies in the classroom and described movies "as an appetizing device for achieving other ends in school".

The reason for choosing courtroom drama, in particular, to examine movies in further detail largely depends on the fact that the interaction of courtroom discourse is usually considered “the closest approximation to everyday speech of all public legal discourses” (Williams, 2005, p. 24); which is an attribute which favors comparisons with previous research (Forchini, 2012a). Besides, the fact that the American adversary procedure is reminiscent of ancient drama makes legal drama an interesting field to investigate from a linguistic point of view, in that it makes language the main actor of the legal event. Moreover, keeping in mind the above-mentioned motivating factor, Courtroom drama is one of the most popular American movie genres which is widely appreciated by the general audience (Kuzina, 2001), so much so that “due to the world wide dominance of American films, viewers in countries with very different legal traditions think their trials follow the United States movie pattern” (Rafter, 2001, p. 24, see also Machura & Ulbrich, 2001). Considering all this, movies can then become a productive source for teaching / learning specialized features of legal language also, for example, to illustrate the rationale for law, and/or to foster critical thinking and analytical skills (see also Ennis, 1989).

Methodologically, real trials are compared to movie trials using corpus-driven criteria (Francis, 1993; Biber, 2009) and Biber’s (1988) Multi-Dimensional Analysis. Data are retrieved from a new sub-corpus of the American Movie Corpus (henceforth AMC, Forchini, 2012a), namely, the American Movie-Trial Corpus, and from the American Real-Trial Corpus (henceforth ARTC, built for the present analysis). Findings show a striking similarity between the two corpora which strongly contrasts the view that movie language is artificial and, thus, not likely to represent spoken language (Gregory and Carroll, 1978; Sinclair, 2004a). The data, indeed, demonstrate that also specialized movie language resembles real language both in terms of textuality and the linguistic items involved in creating such textuality. In particular, the AMTC and the ARTC appear to have four out five Biberian Dimensions in common by displaying involved production (cf. D1), situation-dependent reference (cf. D3) non overt expression of persuasion (cf. D4), and abstract information (cf. D5). Furthermore, although they differ as regards Dimension 2 (i.e. narrative vs. non-narrative concerns), since movie trials are non-narrative and real trials narrative, it is demonstrated that this distinction is only apparent: although real trials have a higher distribution of past tenses, both the corpora have a rather high occurrences of present tenses, which are frequent in spoken interaction. This also explains why the two
trials types also result in having a common most significant dimension, i.e. D1: as textual types, they are both strongly characterized by those linguistic features which contribute to an affective, fragmented, interactional, and generalized context (Biber, 1988), rather than other linguistic features marking other dimensions, which are also typical of spoken language and movie language. It is concluded that the data, by supporting the findings of previous research on the similarity of movie and face-to-face conversation (Forchini, 2009b, 2010b, 2011, 2012a), do, on the one hand, confirm the textual type of movie language, and, on the other hand, re-confute the claim that movies have “a very limited value in a general corpus, because they are ‘considered’ language, written to simulate speech in artificial settings” and thus “are not likely to be representative of the general usage of conversation” (Sinclair, 2004a, p. 80). The present findings also disprove the claim that the cinematic portrayal of the American legal system is far removed from legal reality (Machura & Ulbrich, 2001, p. 118), at least, as far as language is concerned. Needless to say, they also add further value to the role of corpora in teaching, which is often emphasized by numerous authoritative linguists (Hunston, 2002; Mauranen, 2004; Sinclair, 2004b; Reppen, 2010; inter alia). Given the similarity of real and movie language, indeed, the main application which derives from the present study is that both teachers and students could benefit from using movie corpora to either teach or learn not only the features characterizing spoken discourse (Forchini, 2012a, 2012b, 2013b), but also those specialized features characterizing legal discourse.

References


Forchini, P. (2013b). The teaching applicability of movies and the strength of Multi-Dimensional Analysis (MDA). In A. C. Murphy & M. Ulrych (Eds.), *Perspectives on Spoken Discourse* (pp. 81-110). Milano: EduCatt.


Language Education.
The British Academic Written English (BAWE) corpus contains about 6.5 million words of proficient university student writing, categorised in terms of ‘genre families’ and distributed fairly equally across levels of study and disciplinary groupings. The corpus has been examined from various perspectives, including multidimensional analysis (MDA) (Nesi & Gardner 2012) using Biber’s 1988 dimensions. This analysis revealed clear distinctions between texts produced by first, second and final year undergraduates and Masters students, and between texts belonging to different disciplinary groupings. In terms of stance, epistemic modality was the most discriminating feature, being considered a tool for persuasion on Dimension 4. Published academic prose from the Lancaster-Oslo-Bergen (LOB) Corpus had a neutral score on Dimension 4 in Biber’s own 1988 study, but scores for BAWE assignments were negative on Dimension 4, and became increasingly so across levels of study (Table 1). Scores for Arts and Humanities assignments were particularly low (Table 2).

Biber’s 1988 dimensions still have validity, because the results from studies which employ them can easily be compared. For example, the 1988 dimension 4 scores for BAWE and Biber’s LOB subcorpus can be taken as an interesting indicator of differences between student and expert writing. Student writing below PhD level appears to have fewer overtly persuasive features than expert writing, perhaps because undergraduate and Masters students tend to discuss the work of others rather than presenting their own original research.

Stance theory has developed considerably since 1988, and features have now been added to Biber’s tagset, enabling clearer distinctions to be made between registers which differ predominantly in terms of stance. Biber (2006) uses this tagset as the basis for a detailed linguistic description of registers encountered by university students, such as classroom teaching, coursepacks and institutional writing, but does not include any analysis of texts (such as assignments) produced by students themselves. A new multidimensional analysis of BAWE (BAWE2016) has built on this work, using the extended tagset to create new dimensions specific to university student writing. These bespoke dimensions distinguish between BAWE texts grouped by discipline and genre family, without reference to registers in other corpora.

Following MDA methodology, 39 linguistic features were retained and four new factors were identified (Gardner, Nesi & Biber, under review), to be treated as dimensions along which the BAWE registers could be situated. Two of the four new dimensions depend strongly on stance features. The variation along each dimension was then

<table>
<thead>
<tr>
<th>Level</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st year undergraduate</td>
<td>-1.4</td>
</tr>
<tr>
<td>2nd year undergraduate</td>
<td>-1.4</td>
</tr>
<tr>
<td>3rd year undergraduate</td>
<td>-1.5</td>
</tr>
<tr>
<td>Masters</td>
<td>-2.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Disciplinary Grouping</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical Sciences</td>
<td>-1.2</td>
</tr>
<tr>
<td>Social Sciences</td>
<td>-1.3</td>
</tr>
<tr>
<td>Life Sciences</td>
<td>-1.5</td>
</tr>
<tr>
<td>Arts and Humanities</td>
<td>-2.3</td>
</tr>
</tbody>
</table>

Table 1: BAWE Dimension 4 scores by level

Table 2: BAWE Dimension 4 scores by disciplinary grouping
interpreted by qualitative means, manually examining high- and low-scoring texts to ascertain the communicative effect created by the clustering of features.

Stance adverbials and stance nouns controlling that-clauses clustered with 3rd person pronouns, proper nouns and communication verbs at the negative pole of Dimension 1. Together these have been interpreted as indicating ‘stance toward the work of others’. Essays tend to have the lowest scores on this dimension, particularly in Arts and Humanities disciplines (Table 3).

<table>
<thead>
<tr>
<th>Discipline</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparative American Studies</td>
<td>-6.78</td>
</tr>
<tr>
<td>History (n=95)</td>
<td>-6.89</td>
</tr>
<tr>
<td>English (n=106)</td>
<td>-7.38</td>
</tr>
<tr>
<td>Classics (n=82)</td>
<td>-10.53</td>
</tr>
<tr>
<td>Philosophy (n=106)</td>
<td>-10.67</td>
</tr>
</tbody>
</table>

Table 3: Disciplines with the lowest scores on Dimension 1

Stance nouns controlling that-clauses were relatively rare in Biber’s university language corpus, with the exception of fact (Biber 2006:112). In texts with low scores on Dimension 1 most of these nouns have an epistemic function, but some fall into the category of ‘Communication’ or ‘Attitude’ nouns (Biber 2006:93). Stance nouns in the 20 lowest-scoring texts are listed in Table 4.

<table>
<thead>
<tr>
<th>Likelihood</th>
<th>Certainty</th>
<th>Communication</th>
<th>Attitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>Belief</td>
<td>9</td>
<td>2 Argument</td>
<td>1 Fear</td>
</tr>
<tr>
<td>Assumption</td>
<td>2</td>
<td>3 Confession</td>
<td>1 Feeling</td>
</tr>
<tr>
<td>Claim</td>
<td>1</td>
<td>30 Explanation</td>
<td>1 Hope</td>
</tr>
<tr>
<td>Hint</td>
<td>1</td>
<td>1 Interpretation</td>
<td>1 Illusion</td>
</tr>
<tr>
<td>Hypothesis</td>
<td>1</td>
<td>1 Intimation</td>
<td>1 Indication</td>
</tr>
<tr>
<td>Idea</td>
<td>11</td>
<td>1 Objection</td>
<td>1 Understanding</td>
</tr>
<tr>
<td>Interpretation</td>
<td>1</td>
<td>1 Proposal</td>
<td>1 View</td>
</tr>
<tr>
<td>Intimation</td>
<td>1</td>
<td>2 Proposition</td>
<td>3</td>
</tr>
<tr>
<td>Notion</td>
<td>3</td>
<td>1 Statement</td>
<td></td>
</tr>
<tr>
<td>Possibility</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Premise</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Presumption</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sign</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Suggestion</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>41</strong></td>
<td><strong>Total 42</strong></td>
<td><strong>Total 20</strong></td>
</tr>
</tbody>
</table>

Table 4: Stance nouns in the 20 lowest-scoring texts on Dimension 1.

Likewise stance adverbials mainly serve as epistemic devices, particularly to indicate certainty. Table 5 lists these items under the categories identified by Biber (2006:92).

<table>
<thead>
<tr>
<th>Likelihood</th>
<th>Certainty</th>
<th>Style</th>
<th>Attitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apparently</td>
<td>3</td>
<td>Actually</td>
<td>9 Importantl y</td>
</tr>
<tr>
<td>Arguably</td>
<td>3</td>
<td>Always</td>
<td>9 Clearly</td>
</tr>
<tr>
<td>Evidently</td>
<td>3</td>
<td>Certainly</td>
<td>15 Frankly</td>
</tr>
<tr>
<td>Kind of</td>
<td>13</td>
<td>Decidedly</td>
<td>2 Fundamenta lly</td>
</tr>
<tr>
<td>Maybe</td>
<td>1</td>
<td>Doubtless</td>
<td>1 Generally</td>
</tr>
<tr>
<td>Perhaps</td>
<td>10</td>
<td>Evidently</td>
<td>3 Mainly</td>
</tr>
</tbody>
</table>
Table 5: Stance adverbials in the 20 lowest-scoring texts on Dimension 1.

At the positive pole of Dimension 2, stance verbs controlling to- and that-clauses occur alongside mental verbs, that-deletions, 1st person pronouns and past tense verbs. Together, these have been interpreted as indicating ‘Personal stance’. Genres that are not typically considered ‘academic’ - reflective writing, legal Problem Questions discussing the implications of everyday situations, and Empathy Writing engaging with non-experts – tend to contain more of these features. Scores are highest in the soft and applied disciplines (Table 6).

Table 6: Disciplines with the highest scores on Dimension 2

Stance verbs in the 20 highest-scoring texts on Dimension 2 are shown in Tables 7 and 8, using the categories identified by Biber (2006:92).

Table 7: Stance verbs controlling that-clauses in the 20 highest-scoring texts on Dimension 2.
At the positive pole of Dimension 2 stance features indicate a different register, more expressive of personal attitudes, desires and efforts. Stance verbs suit texts that are less like conventional university assignments because they appeal to readers at a more emotional level, and are essentially designed to persuade.

These findings indicate some distinct differences in the expression of stance across genres and disciplines, and provide evidence to support the more general claims made in prior studies. For example Hyland has argued that the soft disciplines are “more inclined to explicitly recognise the role of human agency in constructing knowledge” (2002:124) and to take “more involved and personal positions than those in the science and engineering fields” (2015:34).

Epistemic nouns and adverbials are particularly prevalent in soft discipline Essays, because argumentative claims in this register tend to be developed on the basis of certainty and likelihood (see, for example McEnery & Kifle 2002). Stance nouns which control \textit{that} -clauses also enable arguments to unfold, by serving as ‘shell nouns’ (Hunston & Francis 1999) or ‘signalling nouns’ (Flowerdew & Forest 2014) which label and encapsulate information presented elsewhere in the text. Epistemic and attitude adverbials were strongly associated with spoken registers in Biber’s university language corpus (2006:103-4), but in texts at the negative pole of Dimension 1 they are used to boost and hedge, signalling the writer’s direct involvement in the text. Again, this is typical of the soft disciplines where, in contrast to the hard sciences, there is “less control of variables, more diversity of research outcomes, and fewer clear bases for accepting claims” (Hyland 2015:35).

Our study has also shown that there is a distinctive register found in student writing that employs first person pronouns and ‘stance’ \textit{to-} and \textit{that-} clauses. Future studies could investigate the extent to which the personal stance registers of student writing are also

<table>
<thead>
<tr>
<th>Probability/ Cognition/ Perception</th>
<th>Desire/ Intention/ Decision</th>
<th>Speech Act/ Communication</th>
<th>Causation/ Modality/ Effort</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appear 3</td>
<td>Agree 2</td>
<td>Ask 3</td>
<td>Attempt 5</td>
</tr>
<tr>
<td>Believe 1</td>
<td>Choose 9</td>
<td>Be said 1</td>
<td>Enable 22</td>
</tr>
<tr>
<td>Expect 4</td>
<td>Decide 7</td>
<td>Teach 7</td>
<td>Encourage 9</td>
</tr>
<tr>
<td>Find 1</td>
<td>Hope 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Forget 1</td>
<td>Intend 2</td>
<td></td>
<td>Help 31</td>
</tr>
<tr>
<td>Know 1</td>
<td>Like 6</td>
<td></td>
<td>Manage 5</td>
</tr>
<tr>
<td>Learn 8</td>
<td>Mean 3</td>
<td></td>
<td>Persuade 1</td>
</tr>
<tr>
<td>Seem 12</td>
<td>Need 37</td>
<td></td>
<td>Require 4</td>
</tr>
<tr>
<td>Tend 2</td>
<td>Plan 2</td>
<td></td>
<td>Try 29</td>
</tr>
<tr>
<td>Suppose 3</td>
<td>Prefer 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prepare 1</td>
<td>Want 78</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wish 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong> 28</td>
<td><strong>Total</strong> 156</td>
<td><strong>Total</strong> 11</td>
<td><strong>Total</strong> 108</td>
</tr>
</tbody>
</table>

Table 8: Stance verbs controlling \textit{to}-clauses in the 20 highest-scoring texts on Dimension 2.
found in expert writing, perhaps extending beyond research writing to include texts produced in the professions.

References


The use of general academic and discipline-specific corpora for research writing: introducing data-driven learning to PhD students in Hong Kong

Laurence Anthony (Waseda University, Japan), Meilin Chen (City University of Hong Kong, Hong Kong) and John Flowerdew (Lancaster University / University of London, UK)

There is no doubt that the advent of electronic corpora has revolutionised many areas of linguistic research, including applied disciplines such as language teaching, reference publishing, machine translation, and speech recognition. In language teaching, the use of corpora is no longer restricted to researchers and teachers; in the past 20 years or so there has been a growing research trend of investigating the direct use of corpora by language learners, referred to as data-driven learning (DDL, [Johns, 1991]). Indeed, Boulton and Cobb’s (2017) comprehensive review identified more than 100 studies reporting on DDL endeavours to date. Notable DDL courses reported in the field of academic writing include Lee and Swales’s (2006) DDL course for a small group of doctoral students at the University of Michigan; the corpus-assisted academic writing course run by Maggie Charles (2007, 2010, 2012, 2015) at Oxford University; similar courses run by Viviana Cortes (2014) for several years at Georgia University; and a course run by Lynne Flowerdew (2015) at the Hong Kong University of Science and Technology. Most of these interventions, however, were small-scale operations, often in experimental conditions (Boulton & Wilhelm, 2006; Leńko-Szymańska & Boulton, 2015) and the DDL approach has not been widely disseminated in mainstream language pedagogy. As Leech (1997: 2) pointed out nearly 20 years ago, the “trickle down” effect from research to teaching may be slower than expected.

This paper reports on a territory-wide project that aimed to disseminate the DDL approach among postgraduate research students in Hong Kong. A half-day workshop was delivered more than 20 times at six of the eight public universities in Hong Kong. In total, nearly 500 students attended the workshop, accounting for 6.7% of the research degree community in Hong Kong. The workshop, which consisted of three parts, introduced different types of corpora to the students. In Part 1, students learned how to use the academic component of the BNCweb corpus to polish their writing at both lexico-grammatical and discourse levels. During Part 2, they were introduced to a discipline- and section-specific corpus (including seven sub-corpora) of research articles with AntConc (Laurence, 2016). Adopting the genre- and corpus-based dual approach (Charles, 2007; L. Flowerdew, 2005, 2009, 2015), students analysed the discourse moves of certain sections of research articles and their linguistic realisations. In Part 3, the final part, students went through the process of creating their own personal corpora under the teacher’s guidance. Selected activities from the workshop will be demonstrated during the presentation.

Results from the post-workshop survey show that the great majority of the students were not familiar with corpora or DDL before attending the workshops, and that they found the workshop to be of high quality and useful for improving research writing. Students indicated that they would continue to use corpus in research writing after the workshop. Their critical suggestions, on the other hand, addressed the intensity of the workshop. The findings show that intensive introductory workshops
can be an effective way of teaching PhD candidates to learn to write for publication purposes independently using the DDL approach, there is a trade-off between highly concentrated work and more time spent on more exemplification and practice. The introduction of this approach among research candidates, nevertheless, is a highly valuable and feasible alternative or additional solution for institutions in which writing support is inadequate for PhD students.

References


The Development and Evaluation of a Corpus-based Spanish Collocation Error Detection and Revision Suggestion Tool
Lu Hui-Chuan (NCKU, Taiwan) and Cheng An Chung (University of Toledo, USA)

The topic of collocation has drawn researchers’ attention for the past thirty years in the lexical area of theoretical and applied linguistics (e.g., Firth, 1957; Nattinger, & DeCarrico, 1992; Sinclair, 1991; Wray, 2000). However, available assisting tools for learning Spanish collocation are much less than those for learning English in the field (e.g., Alonso Ramos, Nishikawa, & Vincze, 2010; Boisson et al. 2013; Bolshakov, & Miranda-Jiménez, 2004; Chen, 2011). This study expanded the functions of two previously constructed corpora and furthered the application of our research products in learning Spanish collocation. It included two primary tasks: (1) the development of a modified Spanish collocation assisted learning tool and (2) the evaluation of the relative effects of the developed assisted tool for learning collocation. The research question of the study was: Do the functions of the Spanish Collocation Error Detection and Revision Suggestion Tool assist learners to improve their learning?

In this study, a corpus-based assisted tool for learning Spanish collocation was developed and evaluated. First, based on the training data compiled in two created corpora (Taiwanese Learners’ Corpus of Spanish and Parallel Corpus of Spanish, English and Chinese) and an extraction tool of Spanish collocation (Spanish Collocation Tool), the developed tool was designed with two main functions: error detection and revision suggestion of Spanish collocation for learning purposes. The database of Spanish collocations was generated with machine learning and through the procedure of data processing, collocation extraction, and manual modification. The programming languages used for the interface of users’ query at runtime are HTML, Perl and PHP with Jquery kit and mysql. This developed assisted tool can detect inappropriate learner uses of Spanish collocation and provide suggestion lists for them to choose to revise the collocation errors.

In the second part of the study, a pretest, a video tutorial, a posttest and a questionnaire were conducted to evaluate the effectiveness of the assisted tool. The relative effectiveness of this tool was evaluated using an experiment to compare learning outcomes and feedback from two groups of Spanish learners (experimental and contrastive groups). Research results showed that our developed tool assisted learning effectively given the progress made from pretest to posttest. Furthermore, the results of satisfaction survey on interface and usefulness demonstrated that most of the participants positively confirmed the achievement of this tool for its effectiveness in assisting in the practice of Spanish collocations.

With respect to the limitation of this research, the training data for our developed tool from the two corpora was relatively small. Therefore, the identification and detection of errors were limited to collocations in a fixed and limited range. Also, the context and the current experiment were conducted within searchable combinations. Larger amount of training data from more varieties of contexts should be included for training in the future in order to strengthen the applicability of the tool.

This study shed light in pedagogical applications of created corpora and in the learning of Spanish collocation with a corpus-based approach in the setting of multilingual acquisition.
References


A web of analogies: key descriptive constructions in modern fiction
Dirk Siepmann (Universität Osnabrück, Germany) and Susanne Dyka (Universität Osnabrück / Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany)

One key feature of modern fiction is the frequent use of descriptive lexis, especially descriptive verbs (verbs that, besides having a core lexical meaning, contain a further semantic element, such as jerk ['movement' + 'sharply arrested'], stiffen, huddle, gasp [cf. Snell-Hornby 1983]) and adverbs (e.g. swiftly, suddenly, briskly). Although most such words are also found in everyday language, the patterns in which they occur in fiction are often specific to that genre; thus, the motion verb jerk, for example, occurs in resultative (depictive) constructions with adjectival and prepositional complements of the type awake, erect, into wakefulness, etc. (Siepmann 2015); similar considerations hold true for verbs of gesture such as nod, which in literary usage frequently occurs in a ‘reaction object construction’ (Levin 1993, Martinez Vazquez 2014), typical collocations being nod (one’s) agreement/approval/assent/understanding/satisfaction/encouragement.

The patterns underlying such recurrent literary uses of language may be described in terms of argument-structure-constructions (Goldberg 2006). An attempt will be made to show how valency-increasing argument-structure-constructions, especially resultative and reaction object constructions, are systematically exploited by novelists to express dynamic action in four major descriptive fields (human behaviour, movement and position, sound, and light) which are related through a complex web of analogies. The use of argument-structure-constructions involving descriptive verbs will be shown to be one of the most salient but as yet under-researched characteristics of fictional style. Comparison will be made with French, where similar constructions are far less complex and varied.

References

Rachel Wyman (King’s College London, UK)

The 2016 U.S. Presidential election was explosive, with both candidates – Hillary Clinton and Donald Trump – mired in a seemingly endless array of scandals. Trump’s win was swiftly followed by an enraged outcry from the liberal mainstream media in which those who voted for him were condemned as racist, xenophobic, sexist, homophobic and unintelligent. Trump countered these claims by Tweeting that it was his debate performances that had resulted in his victory. In these debates, Clinton’s strategy was to continue the democratic establishment narrative that Obama has created over the course of eight years, largely pushing for the same policies and repeating the same arguments. Trump’s strategy was mainly to refuse Clinton’s arguments; he regularly stated that the reality Clinton was creating was absolute fiction. In its place, he offered a new political narrative on America, one in which President Obama had abjectly failed the American populace, regarding both foreign and domestic policy. In place of the economic domestic recovery and increasingly peaceful world Obama now claims as his legacy, Trump argued that the economic recovery has been a farce supported with misleading statistics and that the war on ISIL in the Middle East has been a humiliating blunder, resulting in decreased American global hegemony. This presentation details the results of a corpus and argument reconstruction-based critical discourse analysis of Obama’s political narrative and how the prototypical arguments within it eventually came to be defeated by the emergence of Donald Trump’s conflicting narrative during the 2016 Presidential debates.

This study utilizes a new 5-step research approach which combines corpus linguistics, qualitative coding and argumentation analysis. The data for this study are 413 of Obama’s Weekly Addresses to the Nation, spanning both of his terms, 2009 – 2017, as well as the 2016 Clinton/Trump presidential debate transcripts. In Step 1, a corpus analysis of these combined speeches is undertaken to identify the lexical and semantic patterns in Obama’s language, as well as how he frames key issues. In Step 2, the major argument in each individual speech is isolated and qualitatively codes into five separate functional units with NVivo: goals, claims for action, means-goal, circumstances and values (Fairclough & Fairclough 2012). These coded units are then used to create five distinct corpora, which are then quantitatively and qualitatively analyzed, to illuminate the distinctive characteristics of each (O’Halloran 2011). The resulting data shows the key frames that define each of these functional units as well as how Obama’s arguments on his major topics develop over time. In Step 3, the trajectory of Obama’s arguments on each of these major topics is analyzed using a process of comparative argument reconstruction, to identify arguments that have a high probability of being flawed. These results are used, along with the corpus data from Steps 1 and 2, to select the major arguments of Obama’s Presidency and the argumenta that have been flagged as needing further analysis based on the how they develop over time. In Step 4, these arguments are evaluated using argument reconstruction and critical questioning (Fairclough & Fairclough 2012). This evaluation focuses on assessing both the structure of the
arguments as well as the content of the functional units, particularly how key representations are used to reinforce ideologies. In Step 5, Obama’s arguments and the key representations upon which they are based, are then compared to Clinton’s and Trump’s arguments on the same topics. This comparison shows how the establishment democrat narrative, as developed by Obama and continued by Clinton, works to depict a very different world than the one that Trump portrayed in his debate arguments. The result was an election in which voters had to choose between one of two contradicting versions of reality.

The corpus analysis results show that Obama’s political narrative focuses on a group of eleven major subject areas. Within each he uses a progression of arguments to chart a story of progress on each issue, spanning his eight years in office. Each of these minor narratives works within the greater over-arching Presidential narrative which focuses on America’s fight to recover from the Great Recession, culminating in the nation’s successful economic recovery – the major goal of Obama’s Presidency. By incorporating a strategy of highly repetitive language, frames and arguments, he constructs a complex political narrative which works to create the version of reality that he wants the American public to see. But the argument analysis results show that many of Obama’s major arguments, which form the basis of this narrative, so not stand up well to critical evaluation. It is Obama’s argument’s in reference to Syria, which Clinton adopted and enhanced by involving and attacking Russia, which ultimately led Trump to outperform her in the second and third debates. He did so by rejecting the reality that Obama had worked for eight years to create and in doing so, established a new reality in which Syria, Russia and Iran – represented as villains in Obama and Clinton’s speeches – were on the front lines fighting ISIL. I argue that it was this argument, along with Obama’s inability to convince the populous that an economic recovery had occurred, that led to the shocking election of Donald Trump.

References

Conditionals in Eighteenth and Nineteenth century scientific writing
Luis Puente-Castelo (Universidade da Coruña, Spain)

Science and scientific writing underwent deep transformations starting in the seventeenth century. Escaping from the introverted medieval scholastic paradigm, the scientific register used during the first stages of the process of transformation towards contemporary science was characterised by the narration of real life discoveries or experiments, often in first person narratives in which the position of the author in the scientific community was sufficient proof of veracity of the events narrated. This started to change from the late seventeenth and early eighteenth century, with an evolution towards a higher importance of proof and method.

However, as scientific communities grew bigger and authors started having to persuade their peers of the veracity of their accounts, a parallel phenomenon appeared: scientific writing started to develop a series of features to convey modesty, humility, and politeness (Atkinson 1996, 1999; Hyland 1996, 1998, 2000), such as the use of modal verbs, and expressions of probability, evaluation, or attribution; which are used by authors to avoid controversy and achieve a better reception of their work. Conditional structures play an important role among these strategies. Conditionals are frequently used in scientific writing to advance the arguments put forwards by the author, expressing the relationship between phenomena and its consequences, as well as to speculate on possible outcomes and to state hypotheses and proposals. However, conditionals also perform a number of persuasive roles: they are used to define shared assumptions and thus narrow the scope of claims, to tone down claims and mitigate their force, or to directly express politeness or uncertainty, among others (Carter-Thomas & Rowley-Jolivet 2008, Warchal 2010).

This paper’s aim is to study the uses and functions of conditionals in eighteenth and nineteenth century scientific writing, and to show how conditional uses reflect the general evolution of scientific writing during the period, as well as the differences between the different disciplines in the study.

This research has been carried out using the Coruña Corpus of English Scientific Writing (henceforth Coruña Corpus or CC) (Crespo & Moskowich, 2010; Moskowich 2011). The Coruña Corpus is a corpus of late Modern English scientific writing covering the period between 1700 and 1900. It is divided in a series of twin subcorpora, all of them sharing the same structure and principles of compilation, and one for each different discipline. Each of these subcorpora contains forty samples approximately 10,000 words long, at a rate of two samples per decade, so that each subcorpus contains c.400,000 words. Three full subcorpora have been used in this study: CETA (the subcorpus on Astronomy), CEPhiT (the subcorpus on Philosophy), and CELiST (the subcorpus on Life Sciences). All of them together add up to c.1,200,000 words.

These corpora have been searched with the help of the Coruña Corpus Tool (Moskowich & Parapar 2008) for selected conditionals particles, obtained adapting Quirk et al.’s (1985) and Declerck & Reed’s (2001) classifications so that they reflect the uses in the period. Each of the results obtained has been manually
disambiguated so as to eliminate non-conditional uses, such as interrogative if uses, leaving 3,735 occurrences in the corpus.

Each of these occurrences has then been recorded with sufficient context, and classified according to nine different parameters: five of these are socio-historical or extra-linguistic (the period of the text, its discipline, its genre, the sex of the author, and their geographical origin) and four are linguistic (the type of conditional particle, the order of the constituents, the verb-form combinations, and the function conditionals play in discourse; this latter parameter being classified with the help of a new, specific typology). The results show that there are important differences both in the number and the types of conditional used.

In what has to do with the general frequency of use of conditionals, there are differences in their use across disciplines during the eighteenth century, with a higher use in philosophy (particularly among women) and a lower use in both astronomy and life sciences. During the nineteenth century, these differences became less important. The differences found regarding the different types of conditionals are interpreted as a reflection of the different stages of popularisation of the new paradigms of science and scientific writing in each of the disciplines. Thus, during the eighteenth century, the high number of conditional uses in philosophy, as well as the general greater variability, the starker differences among male and female philosophers, and the greater use of conditionals other than if point to a scenario in which the influence of the old scholastic paradigm is still present in philosophy texts. This contrasts with the lower and more standardised use of conditionals in both astronomy and life sciences during the period, although there are also some particularities in these disciplines, such as the high number of non-mitigating conditionals in astronomy texts, which shows the influence of Newton's preference for the mathematical expression of reality. During the nineteenth century, however, the differences among disciplines are less important, thus reflecting the evolution of philosophy towards the practices of the new scientific paradigm, and a more standardised scientific register in which the differences between disciplines are less important.

There are as well some important differences according to the parameter of sex, as women tend to avoid the most blatant types of mitigating conditionals, preferring to mitigate their claims with the use of modality instead. This may be related to the pressures on women scientists, which could have made them select the less salient types of mitigation.

References


University students need to do a great deal of reading. For second language students, this is a daunting prospect due to the vocabulary load. As readers, they need to understand around 95-98% of the words ('tokens') in a text in order to comprehend what they read, and a high proportion of this lexis is discipline specific (Hyland & Tse, 2007; Durrant, 2014).

Reports vary as to the actual amount of discipline-specific vocabulary required, depending not only on the subject but also on the methodology used - such as whether high frequency general English vocabulary (e.g. the new-GSL - Brezina & Gablasova, 2013) and general academic vocabulary (e.g. the Academic Vocabulary List - Gardner & Davies, 2014) are included or excluded. For instance, should the words law and court, which are both high frequency items generally understood by a non-expert, be excluded from a list of words for second language law students? It is probably true that for some students, such as foundation students, they should be included, whereas for doctoral students they would be known. Therefore, a word list should not only have a clearly defined purpose (Nation, 2016), but also be clearly interpretable by the user, who is unlikely to have a background in Corpus Linguistics or be an expert in word lists. Furthermore, word lists tend to be single words, or collocations, or multiword units. However, as a teacher, as well as a researcher, I am interested in all three. For example, the phrase in (the) light of is frequently used to connect ideas in text; and although it may be considered sub-technical, general academic vocabulary, it actually has a very specific usage in legal texts. Owing to the specific nature of the lexis under investigation, I am also interested in what the subject experts have to say about the vocabulary their students need.

The innovative idea behind the Discipline-Specific Vocabulary Core (DSVC) is that it will consist of separate 'building blocks'. In this way the various lexical items can more easily be incorporated into pedagogical resources. This poster will discuss the methodology being trialled to identify the vocabulary items and also illustrate the different DSVC building blocks with data from the DSVC International Law corpus, which is under construction.

Existing law corpora (e.g. Williams, 2007; Marín & Rea, 2012) are compiled from primary sources, such as law reports. The DSVC International Law corpus consists of a wide variety of both primary and secondary genres from postgraduate Law reading lists and is therefore much more representative of learner needs. While the focus of my research is on the vocabulary these Law students need for reading, I also maintain the view that when a student writes an essay, their main lexical resource is their reading. Thus the DSVC International Law corpus is intended to be used to form the basis of a range of teaching materials and other resources, including opportunities for hands-on concordancing.
References


Discipline, discourse and new viewpoints. A Report on the Coruña Corpus

Isabel Moskowich, Begoña Crespo, Luis Puente-Castelo and Leida
Maria Monaco (University of A Coruña, Spain)

The Coruña Corpus of English Scientific Writing (henceforth, Coruña Corpus or CC) is a corpus of late Modern English scientific writing, covering the period between 1700 and 1900.

When the project started back in 2003, the Coruña Corpus was influenced by a growing interest and gradual increase in studies in both the vernacularisation of Science (Crossgrove 2000) and genre conventions (Taalvitsainen 2004). With this idea in mind, the Coruña Corpus has been designed as a tool for the study of language change in English scientific writing in general and within each different scientific discipline in particular.

The CC has been conceived of as a collection of several twin corpora, one for each different discipline, but all sharing the same structure, organisation, mark-up, and principles of compilation. Each subcorpus contains samples from a given discipline ( delimited by adapting UNESCO’s classification of Sciences to the reality of the period and with an inclusive view) at a rate of two c.10,000-word samples per decade, thus adding up some 400,000 words per subcorpus. Each sample is encoded in XML following TEI conventions, respecting the original layout and spelling variables as much as it is feasible. Each sample is accompanied by a metadata file containing information about the author of the text as well as about the text itself.

The Coruña Corpus is exploited with a specifically designed information retrieval application, the Coruña Corpus Tool (CCT), which works as most concordancers but offers the possibility of searching old characters, spelling variants, and tags; as well as selecting samples on account of metadata parameters.

As of now, the subcorpora on Astronomy (Moskowich & Crespo 2012) and Philosophy (Moskowich et al. 2016) have been finished and published, and the subcorpora on Life Sciences, History and Chemistry are in several stages of beta development, with compilation of the samples having finished but work on metadata still in process. Work is now underway in the compilation of the subcorpus on language and linguistics. The fact that disciplines pertaining to the traditionally labelled hard and soft sciences are compiled allows also comparative studies regarding the evolution of different discursive trends.

This poster presents the latest developments in the project, with special attention being paid to the corpora on history, chemistry, and linguistics.

References

Crossgrove, W. (2000). The Vernacularization of Science, Medicine, and


Saying Whatever It Takes: Creating and Analyzing Corpora from US Presidential Debate Transcripts
Leo Vrana (University of Konstanz, Germany) and Gerold Schneider (University of Konstanz, Germany/University of Zurich, Switzerland)

We first describe the creation of a corpus of American presidential debates from the American presidency project. We then use the corpus to present a stylistic analysis of presidential candidates from 2000 to 2016. A range of stylistic measures, including vocabulary richness, language complexity, and readability measures is applied. We aim to contribute to the current debate on the complexity of American presidential rhetoric and the role of the register of spoken language, by furnishing empirical data.

1. Introduction

In the American news media, United States Presidential elections are dissected and examined with a fervor normally reserved for sports championships and natural disasters. This is not surprising given what is at stake. Our contribution focuses on the creation of a corpus of Democratic and Republican presidential nominees’ speech during these debates dating back to the year 2000, and presents stylistic results gleaned from this corpus, including measures of vocabulary richness, language complexity and readability. Further, we compare these results against another corpus of spoken American English in order to possibly uncover any further insights into the differences between the speech of politicians and everyday citizens.

While word-choice and topics vary depending on the agenda of the candidate and on current issues, stylistic features can be varied freely to convey a message tailored to targeted voters. Lim (2008) has claimed that American presidential rhetoric has become considerably less complex over time, a trend that created lively discussions during the election of George W. Bush and even more so with Trump's election. Simpler language in speeches may signal an attempt to address a broader, less educated audience and to satisfy the demands of “sound bite” journalism (Hallin, 1992), but may also be a way to address the demands of the spoken genre, where high complexity increases the risk of ambiguity. This paper also aims to provide empirical data to complement these analyses.

2. The Source and Corpus Creation

The American Presidency Project is a non-profit undertaking hosted online by the University of California at Santa Barbara which archives documents related to United States presidents for public use, including transcriptions of the presidential debates, accessible at http://www.presidency.ucsb.edu/. The first goal of this project was to

---

1 Files related to this project can be found at https://github.com/LeoVrana/PresidentialDebates.
take these transcriptions and compile utterances of each candidate into a corpus for that candidate.

In order to accomplish this, we took the page source for each transcript, and used regular expressions and capturing groups to automatically create a corpus for each candidate. We included a manual validation step to ensure that the data was processed correctly.

Fig. 1: Flow Chart

3. Methods and Results

After ensuring that the files were accurate, we performed the analyses below. We also include figures from the Santa Barbara Corpus of Spoken English (SBC) as a reference where possible (Dubois, Chafe, Meyer & Thompson, 2000-2005). Statistical significance tests were performed to compare politicians relative to each other – values from the SBC were not considered.

3.1 Word-Level Analyses

The mean word-length shows that candidates were not far apart from each other, with the statistically significant exception of Donald Trump ($z = -2.89, p < 0.01$). The average word length from Santa Barbara Corpus was similarly shorter than the other politicians, perhaps partly due to the transcription method. The related count of average syllables per word shows very similar results. Trump was found significantly differ from other candidates here as well ($z = -2.8, p < 0.01$).
3.2 Vocabulary Richness

Measures of vocabulary richness are typically based on type-token ratios. Since tests of vocabulary richness are affected by the size of the corpus, any analysis must account for this (Malvern et al. 2004, Lu 2014: 82). We used the Mean-Segmental TTR (MSTTR), which calculates TTRs for segments of text, and then finds the mean of those TTRs (Lu 2014: 82). A smaller TTR indicates less varied speech.

The measurement shows a large variance, with Trump’s vocabulary being significantly less varied than other politicians ($z = -2.84$, $p < 0.01$), as well as the SBC.
3.3 Bigram Collocations

Analyses of collocations, in addition to linguistic multi-word entities, often bring up key concepts, particularly when analyzing monotopical texts. Collocation research has a long tradition, see Pecina (2009) or Evert (2009). We use the measures Observed over Expected (O/E) and its variant $O^2/E$, which are simple to calculate and interpret, and have a tendency to over-report collocations consisting of rare, i.e. content, words, thus giving insights into candidates' agendas and core interests.

We list the top three bigrams for each candidate below along with the campaign year, sorted by $O^2/E$ (last column).

![Bigram Table]

Table 1: Top 3 collocations per campaign

3.4 Readability

Readability measures partly depend on sentence length. Fortunately, the debate transcriptions added punctuation which made this analysis possible. We determine the ease to which a text can be understood using the *Lingua::EN::Fathom* module by Kim Ryan (available at cpan.org). One of the offered measures is the percentage of “Complex Words,” where a word is considered “complex” if it contains three or more syllables.
One reason for the radical difference for SBC is that the transcriptions in that corpus include hesitation words such as "uh". Trump was found to differ significantly from other candidates ($z = -3.16$, $p = 0.001$). Another measure of readability is words per sentence:

No candidate varied significantly from the rest, but it is interesting to note that with the exception of John Kerry in 2004, Democratic candidates spoke with longer sentences and Republican candidates spoke with shorter sentences. We could not include a comparison to the Santa Barbara Corpus, as the transcription did not include punctuation.

The Flesch readability analysis equation returns a score from 0 to 100, where 100 represents a very easily readable text, and 60-70 is the ideal range (Flesch, 2016). Trump is measured here as being significantly more readable than the other candidates ($z = 2.74$, $p = 0.006$).
While such measures cannot check for grammaticality, they summarize the relative difficulty of understanding each candidate. (Ryan, 2016)

3.5 Surprisal

Readability measures consider words in isolation, but they do not take word sequences into consideration. Psycholinguistic research has shown that routinized sequences are an essential component for ease of readability (Conklin & Schmitt 2012).

We apply surprisal, an information-theoretic measure of the surprise of the continuation of word sequences (Levy and Jaeger 2007). Bigram surprisal is defined as follows:

$$2\text{-gram surprisal} = \log \frac{1}{p(w_1)} + \log \frac{1}{p(w_2|w_1)}$$

Surprisal is the logarithmic version of the probability seeing word $w_1$ linearly combined with the probability of the transition to the next word, $w_2$. The probability $p(w_1)$, is context-independent, while the transitions, e.g. $p(w_2|w_1)$ express predictability in the context. Surprisal is an information theoretic measure; it measures how many bits of information the conversation contains (Shannon 1951): the more expected and thus probable a word is in its context, the less information it carries.
4. Conclusions and Outlook

Our analyses confirmed basic intuitions about speaking patterns of certain candidates, and the concordances and bigrams were interesting reminders of the topics of earlier elections. Differences between transcription methods of the SBC and the presidential debates may be partly responsible for SBC statistics being generally lower than the other politicians, but it must be noted that its statistics were mostly close to Trump’s, and both of those were often quite different from other presidential candidates. This may be seen as quantitative evidence supporting the general public’s perception of Trump as different from all other candidates (irrespective of the content of his speech, which was outside the scope of this paper), and a candidate who sounded more like a “regular” person.

Opportunities for further research could include sentiment analysis, or analysis of metrics such as audience applause or laughter associated with each candidate.

References


Across discipline corpus-based study of logical relation verb patterns in the research article
Min Dong (Beihang University, China)

1. Introduction

A common pattern in abstract or technical writing is to present a consequential conjunction recoded by a verb (Jones, 2010). The notion of PATTERN refers to all the words and structures that are regularly associated with a word and contribute to its meaning (Francis & Hunston, 2000). This paper attempts to investigate disciplinary variation of logical relation verb patterns and associated meanings of causal relation.

2. JC-JDEST

The corpus used is Journal Corpus of the second generation of JiaoDa English for Science and Technology (JC-JDEST).

Table 1 Overall statistics of JC-natural science subcorpus

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Numbers</th>
<th>Parameters</th>
<th>Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural sciences</td>
<td>25</td>
<td>Standardized type/token</td>
<td>37.52</td>
</tr>
<tr>
<td>Text samples</td>
<td>2447</td>
<td>Word length</td>
<td>5.03</td>
</tr>
<tr>
<td>Average text size</td>
<td>1343.17</td>
<td>Sentences</td>
<td>14,9417</td>
</tr>
<tr>
<td>(words)</td>
<td></td>
<td>Sentence length</td>
<td>22.00</td>
</tr>
<tr>
<td>Tokens</td>
<td>340,025</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Types</td>
<td>6,2424</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type/token ratio</td>
<td>1.90</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2 Overall statistics of JC-social science subcorpus

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Numbers</th>
<th>Parameters</th>
<th>Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social sciences</td>
<td>15</td>
<td>Standardized type/token</td>
<td>41.23</td>
</tr>
<tr>
<td>Text samples</td>
<td>683</td>
<td>Word length</td>
<td>5.17</td>
</tr>
<tr>
<td>Average text size</td>
<td>5565.22</td>
<td>Sentences</td>
<td>15,7648</td>
</tr>
<tr>
<td>(words)</td>
<td></td>
<td>Sentence length</td>
<td>24.11</td>
</tr>
<tr>
<td>Tokens</td>
<td>388,1220</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Types</td>
<td>7,8922</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type/token ratio</td>
<td>2.08</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
3. Logical relation verb pattern

In line with ‘favourite clause type’ of English scientific or academic writing, we define logical relation verb pattern as one consisting of two nominal phrases or clauses plus one verbal phrase. In syntactic terms, the two nominal phrases or clauses, whether they are de-verbal nouns, de-adjectival nouns, abstract nouns, embedded finite clauses or embedded non-finite clauses, realize Subject, Object, or Transitive Complement (Fang, 2007), and semantically, they realize two types of activities, that is, argument-oriented activity and event-oriented activity (Gledhill, 1996). The verbal phrase falls into monotransitives or transitives (Fang, 2007) and semantically construe four types of causal relation (Crombie, 1985): 1) Reason–Result construed by two event-oriented activities of causation; 2) Means–Result construed by one or two argument-oriented activities; 3) Condition-Consequence construed by two event-oriented activities of contingency; 4) Grounds-Conclusion construed by two argument-oriented activities. For example,

1) The logical *advances* in computer hardware design during the last few years, *has enabled* the *calculation* of viscous flows at high Reynolds numbers to be made.
2) More detailed *studies* of larger samples of galaxies *ought to allow* more clear-cut *conclusions* in this important field.
3) *How copy cataloging is affected in the realm of classification depends* to some extent *on whether the library uses Dewey Decimal Classification.*

4. Results and Discussion

The results are shown from table 3 to table 7.

Table 3 disciplinary distribution of causal relation patterns of the 8 logical relation verbs

<table>
<thead>
<tr>
<th>Verb</th>
<th>Frequency and Percentage of causal relation patterns in natural science</th>
<th>Frequency and Percentage of causal relation patterns in social science</th>
<th>Significant disciplinary difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALLOW</td>
<td>278 278/1858=14.96%</td>
<td>133 133/1525=8.72%</td>
<td>o.000 (&lt;0.05) sig</td>
</tr>
<tr>
<td>BASE</td>
<td>218 218/3337=6.53%</td>
<td>169 169/3712=4.55%</td>
<td>0.000 (&lt;0.05) sig</td>
</tr>
<tr>
<td>DEPEND</td>
<td>133 133/2130=6.24%</td>
<td>131 131/1348=9.72%</td>
<td>0.000 (&lt;0.05) sig</td>
</tr>
<tr>
<td>ENABLE</td>
<td>120 120/422=28.44%</td>
<td>116 116/463=25.05%</td>
<td>0.256 (&gt;0.05) insig</td>
</tr>
<tr>
<td>INVOLVE</td>
<td>159 159/1436=11.07%</td>
<td>155 155/1942=7.98%</td>
<td>0.002 (&lt;0.05) sig</td>
</tr>
<tr>
<td>MEAN</td>
<td>290 290/2689=10.78%</td>
<td>340 340/5125=6.63%</td>
<td>0.000 (&lt;0.05) sig</td>
</tr>
<tr>
<td></td>
<td>Most frequent causal relation patterns in natural science</td>
<td>Most frequent causal relation patterns in social science</td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>----------------------------------------------------------</td>
<td>--------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td><strong>ALLOW</strong></td>
<td>n V n 36.69%</td>
<td>n V n 48.87%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>n V n to-inf 27.33%</td>
<td>n V n to-inf 24.81%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>this V n to-inf 12.59%</td>
<td>this V n 9.02%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>this V n 10.79%</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>BASE</strong></td>
<td>n V n 81.19%</td>
<td>n V n 76.92%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>n V -ing 11.01%</td>
<td>this V n 5.33%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>this V n 4.13%</td>
<td>n V -ing 5.33%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-ing V n 5.33%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>this V -ing 2.37%</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>DEPEND</strong></td>
<td>n V n 59.40%</td>
<td>n V n 57.25%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>n V wh-cl 9.77%</td>
<td>n V wh-cl 9.16%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>wh-cl V 5.26%</td>
<td>wh-cl V wh-cl 8.40%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>wh-cl V n 5.26%</td>
<td>wh-cl V n 3.82%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>this V wh-cl 2.26%</td>
<td>this V wh-cl 3.82%</td>
<td></td>
</tr>
<tr>
<td><strong>ENABLE</strong></td>
<td>n V n to-inf 40.00%</td>
<td>n V n to-inf 43.97%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>this V n to-inf 20.83%</td>
<td>n V n 16.38%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>n V n 15.00%</td>
<td>this V n to-inf 14.66%</td>
<td></td>
</tr>
<tr>
<td><strong>INVOLVE</strong></td>
<td>n V n 47.17%</td>
<td>n V n 46.45%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>n V -ing 25.79%</td>
<td>n V -ing 18.06%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>this V n 12.58%</td>
<td>this V n 10.32%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>this V -ing 8.18%</td>
<td>this V -ing 6.45%</td>
<td></td>
</tr>
<tr>
<td><strong>MEAN</strong></td>
<td>this V that-cl 62.76%</td>
<td>this V that-cl 47.65%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>n V that-cl 16.90%</td>
<td>n V that-cl 23.82%</td>
<td></td>
</tr>
<tr>
<td><strong>RESULT</strong></td>
<td>n V n 65.22%</td>
<td>n V n 57.66%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>this V n 16.85%</td>
<td>n V -ing 15.33%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>this V n 10.22%</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>SHOW</strong></td>
<td>n V that-cl 73.54%</td>
<td>n V that-cl 82.11%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>n V n 18.78%</td>
<td>n V n 11.58%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>this V that-cl 4.23%</td>
<td>this V that-cl 3.68%</td>
<td></td>
</tr>
</tbody>
</table>
Table 5 Accumulative frequencies of the 8 verbs’ most frequent patterns of argument-oriented activities

<table>
<thead>
<tr>
<th>Verb</th>
<th>Natural Science</th>
<th>Social Science</th>
<th>Significant difference between the two disciplines</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALLOW</td>
<td>n V n</td>
<td>117</td>
<td>n V n</td>
</tr>
<tr>
<td></td>
<td>n V n to-inf</td>
<td>175</td>
<td>n V n to-inf</td>
</tr>
<tr>
<td>BASE</td>
<td>n V n</td>
<td>155</td>
<td>n V n</td>
</tr>
<tr>
<td></td>
<td>n V -ing</td>
<td>155</td>
<td>n V -ing</td>
</tr>
<tr>
<td>DEPEND</td>
<td>n V n</td>
<td>49</td>
<td>n V n</td>
</tr>
<tr>
<td></td>
<td>wh-cl V wh-cl</td>
<td>7</td>
<td>wh-cl V wh-cl</td>
</tr>
<tr>
<td></td>
<td>wh-cl V n</td>
<td>7</td>
<td>wh-cl V n</td>
</tr>
<tr>
<td></td>
<td>this V wh-cl</td>
<td>3</td>
<td>this V wh-cl</td>
</tr>
<tr>
<td>ENABLE</td>
<td>n V n to-inf</td>
<td>64</td>
<td>n V n to-inf</td>
</tr>
<tr>
<td></td>
<td>this V n to-inf</td>
<td>64</td>
<td>this V n to-inf</td>
</tr>
<tr>
<td>INVOLVE</td>
<td>n V n</td>
<td>74</td>
<td>n V n</td>
</tr>
<tr>
<td></td>
<td>n V -ing</td>
<td>74</td>
<td>n V -ing</td>
</tr>
<tr>
<td></td>
<td>this V n</td>
<td>20</td>
<td>this V n</td>
</tr>
<tr>
<td></td>
<td>this V -ing</td>
<td>20</td>
<td>this V -ing</td>
</tr>
<tr>
<td>MEAN</td>
<td>n V that-cl</td>
<td>6</td>
<td>n V that-cl</td>
</tr>
<tr>
<td></td>
<td>this V that-cl</td>
<td>225</td>
<td>this V that-cl</td>
</tr>
<tr>
<td>RESULT</td>
<td>n V n</td>
<td>92</td>
<td>n V n</td>
</tr>
<tr>
<td></td>
<td>this V n</td>
<td>92</td>
<td>this V n</td>
</tr>
<tr>
<td>SHOW</td>
<td>n V that-cl</td>
<td>336</td>
<td>n V that-cl</td>
</tr>
<tr>
<td></td>
<td>this V that-cl</td>
<td>336</td>
<td>this V that-cl</td>
</tr>
<tr>
<td></td>
<td>Total Percentage</td>
<td>1237/1944=63.63%</td>
<td>840/1371 =61.27%</td>
</tr>
</tbody>
</table>

Table 6 Top 10 argument-oriented nouns in the 8 verbs’ most frequent patterns (lemmatized)

<table>
<thead>
<tr>
<th>Verb</th>
<th>Natural Science</th>
<th>Social Science</th>
</tr>
</thead>
<tbody>
<tr>
<td>STUDY</td>
<td>102</td>
<td>STUDY</td>
</tr>
<tr>
<td>ANALYSIS</td>
<td>46</td>
<td>ANALYSIS</td>
</tr>
<tr>
<td>TEST</td>
<td>37</td>
<td>RESEARCH</td>
</tr>
<tr>
<td>METHOD</td>
<td>28</td>
<td>SURVEY</td>
</tr>
<tr>
<td>Discipline</td>
<td>Frequency and Percentage of the Four Types of Causal Relation Construed in Natural Science</td>
<td>Frequency and Percentage of the Four Types of Causal Relation Construed in Social Science</td>
</tr>
<tr>
<td>------------</td>
<td>------------------------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td><strong>ALLOW</strong></td>
<td>Means-Result 18/4 66.19%</td>
<td>Means-Result 92/69.17%</td>
</tr>
<tr>
<td></td>
<td>Condition-Consequence 72/25.90%</td>
<td>Condition-Consequence 26/19.55%</td>
</tr>
<tr>
<td></td>
<td>Reason-Result 22/</td>
<td>Reason-Result 15/</td>
</tr>
<tr>
<td><strong>BASE</strong></td>
<td>Means-Result 14/1 64.68%</td>
<td>Means-Result 87/51.48%</td>
</tr>
<tr>
<td></td>
<td>Condition-Consequence 45/20.64%</td>
<td>Condition-Consequence 43/25.44%</td>
</tr>
<tr>
<td></td>
<td>Grounds-Conclusion 23/</td>
<td>Grounds-Conclusion 23/</td>
</tr>
<tr>
<td></td>
<td>Reason-Result 9/</td>
<td>Reason-Result 16/</td>
</tr>
<tr>
<td><strong>DEPEND</strong></td>
<td>Condition-Consequence 66/49.62%</td>
<td>Condition-Consequence 66/50.38%</td>
</tr>
<tr>
<td></td>
<td>Means-Result 57/42.86%</td>
<td>Means-Result 43/32.82%</td>
</tr>
<tr>
<td></td>
<td>Grounds-Conclusion 10/</td>
<td>Grounds-Conclusion 20/</td>
</tr>
<tr>
<td></td>
<td>Reason-Result 2/</td>
<td>Reason-Result 2/</td>
</tr>
<tr>
<td><strong>ENABLE</strong></td>
<td>Means-Result 73/60.83%</td>
<td>Means-Result 59/50.86%</td>
</tr>
<tr>
<td></td>
<td>Condition-Consequence 42/35.00%</td>
<td>Condition-Consequence 39/33.62%</td>
</tr>
<tr>
<td></td>
<td>Reason-Result 4/</td>
<td>Reason-Result 18/</td>
</tr>
<tr>
<td></td>
<td>Grounds- 1/</td>
<td>Grounds-</td>
</tr>
<tr>
<td></td>
<td>Conclusion</td>
<td>INVOLVE</td>
</tr>
<tr>
<td>---------------</td>
<td>------------</td>
<td>---------</td>
</tr>
<tr>
<td>Condition-Consequence</td>
<td>70</td>
<td>44.03 %</td>
</tr>
<tr>
<td>Grounds-Conclusion</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>MEAN</td>
<td>Grounds-Conclusion</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>Means-Result</td>
<td>73</td>
</tr>
<tr>
<td></td>
<td>Condition-Consequence</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>Reason-Result</td>
<td>1</td>
</tr>
<tr>
<td>RESULT</td>
<td>Reason-Result</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>Means-Result</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Grounds-Conclusion</td>
<td>4</td>
</tr>
<tr>
<td>SHOW</td>
<td>Means-Result</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>Grounds-Conclusion</td>
<td>92</td>
</tr>
<tr>
<td></td>
<td>Condition-Consequence</td>
<td>14</td>
</tr>
<tr>
<td>Total</td>
<td>Means-Result</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td>Condition-Consequence</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>Grounds-Conclusion</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>Reason-Result</td>
<td>29</td>
</tr>
</tbody>
</table>

Here it is shown that 1) most frequent logical relation verb pattern is n V n; 2) unattended *this* (Wulff, Römer & Swales, 2012) span through all of the 8 verb patterns; 3) with frequent nominal categories like n, *this*, *wh-*cl and *that-*cl, which tend to construe argument-oriented activities, the use of logical relation verb patterns is more saliently motivated by inferencing and reasoning in the textual world; 4) they are more typically used to foreground the inherent complexity of
inferential causal relations in natural science.

References


The CorCenCC Crowdsourcing App: A Bespoke Tool for the User-Driven Creation of the National Corpus of Contemporary Welsh
Steven Neale (Cardiff University, UK), Irena Spasić (Cardiff University, UK), Jennifer Needs (Swansea University, UK), Gareth Watkins (Cardiff University, UK), Steve Morris (Swansea University, UK), Tess Fitzpatrick (Swansea University, UK), Lindsay Marshall (Newcastle University, UK) and Dawn Knight (Cardiff University, UK)

Introduction
The CorCenCC project\(^1\) (*Corpws Cenedlaethol Cymraeg Cyfoes* or *National Corpus of Contemporary Welsh* in English; www.corcencc.org) aims to assemble a 10 million-word corpus of the Welsh language across a range of contemporary contexts from spoken, written and e-language sources. In keeping with its contemporary aspect, a key innovation of the project is to facilitate crowdsourced contributions to the corpus, giving Welsh speakers the opportunity to directly involve themselves in the creation of the corpus. This is of vital importance in the Welsh context, in which community pride is strong and for which an open linguistic resource that properly represents the constantly-evolving landscape of contemporary Welsh speakers and the way their language is used is expected to have a wide-reaching impact on the way publishers, policy-makers, the education sector, academic researchers and many more work with Welsh going forward.

This presentation introduces the CorCenCC Crowdsourcing App, a mobile application designed to facilitate direct contribution of spoken language data to the corpus. Spoken language data will comprise 400,000 of the 10 million word corpus (alongside 400,000 word of written data and 200,000 words of electronic language such as blogs and emails), and app users can contribute directly to this number by recording their Welsh-language narratives (Figures 1 and 2), attaching and editing appropriate metadata to describe the recorded conversations, and uploading them for inclusion in the final corpus. The metadata attached to the recorded conversations includes details about where the recording was made, who else was involved in the recording, and tags that future corpus tools will be able to use to search the data in the final corpus.

Sampling and Ethical Considerations
An important consideration of the crowdsourcing app's design is how the data contributed by users correlates to the sampling frame that has been designed for the CorCenCC project, and specifically to the kinds of Welsh users that need to be reflected and accounted for in the data. New users registering to use the app are required to complete a user profile (Figure 3), which elicits a range of information that has been deemed necessary in order to fully represent the Welsh that is being contributed: speaker's year of birth, local authority, regions that have been considered to have influenced their Welsh, context(s) that have influenced the development of their Welsh, and how a speaker would describe their ability to use Welsh, among others. This information about the contributors is vital in enabling

---

\(^1\) Funded by the UK Economic and Social Research Council (ESRC) and Arts and Humanities Research Council (AHRC) as part of the *Corpws Cenedlaethol Cymraeg Cyfoes* (*The National Corpus of Contemporary Welsh*): A community driven approach to linguistic corpus construction project (Grant Number ES/M011348/1).
recorded conversations – and eventual Welsh language data – to be attributed to speakers from various contexts and across different social, economic or geographical backgrounds.

![Figure 1. A list of recordings made using the app.](image1)

![Figure 2. The recording screen](image2)

Being able to reflect this information in contributions made via the crowdsourcing application is of huge importance for CorCenCC, whose user-driven remit has been designed to ensure that contributions are truly representative of the Welsh speaking community. Constructing principled corpora involves ensuring that the data they include is balanced, representative and fully documented, and adequate balance and representation can pose particular challenges to maintain – particularly when data arrives in the form of unplanned, spontaneous and ad-hoc contributions, as in the crowdsourcing context. The information collected about contributors by way of carefully constructed user profiles therefore provides the 'fully documented' aspect of the principled corpus design, allowing for any gaps or discrepancies that might surface in the balance and representation of crowdsourced contributions to be identified and properly addressed.

In keeping with the ethical considerations of corpus creation, ensuring that full permissions are sought from and provided by users of the Crowdsourcing App is a vital consideration. After being shown a full description of the project and how data will be used, we ask participants to allow us permission to use their contributions in CorCenCC and for project-related research activities before they are able to register as a user, and also encourage users to record themselves and anybody else involved in a recording verbally stating that they consent for a conversation to be used before they are able to make an actual recording. Additionally, we fully explain that upon upload for inclusion in the corpus, recordings are appropriately anonymised in the same way as a contribution collected by our linguists in the field would be.
Crowdsourcing and Apps in Applied Linguistics

Crowdsourcing methods – where tasks are outsourced to external contributors (the crowd) – are still at a relatively early stage of adoption in applied linguistics, but their successful application in other research areas – including related fields from other disciplines such as Natural Language Processing – are evidence of their merit and their potential impact. Some examples where crowdsourcing methods have been used in applied linguistics have involved the creation and collection of speech and language data (Callison-Burch and Dredze, 2010; Lane et al., 2010), often using tools designed to facilitate outsourcing tasks to external contributors such as the Amazon Mechanical Turk or by completion of bespoke processes such as an online educational game (McGraw et al., 2010). Crowdsourcing methods have also been used in more defined tasks, such as the transcription of language data for speech recognition (Novotney and Callison-Burch, 2010) and the evaluation of translation quality (Callison-Burch, 2009).

Crowdsourcing of language data by way of bespoke applications is in its early stages of adoption, with the Dialäkt and Voice Äpp built to collect data to help identify and differentiate between different Swiss German dialects in locations around Switzerland being a noteworthy example (Goldman et al., 2014). However, although larger-scale user-driven corpus construction is as yet unexplored, it is particularly viable and appropriate in the Welsh language context, where community pride is of huge importance and there is a passionate interest in sustaining and 'growing' the language. The CorCenCC Crowdsourcing App can offer a vital resource for harnessing this interest and giving Welsh speakers and learners the opportunity to directly contribute to an open, growing resource that is both for them and by them.

Conclusion and Future Work

The Crowdsourcing App has already gone through a technical testing phase and moving forward we are planning a full human-factors evaluation, using tried-and-tested means such as the System Usability Scale (Brooke, 1996) and a qualitative questionnaire to discern how easy the app is for everyday users. The app is currently available for iOS platforms (iPhone and iPad), with an Android version of the app in development that will widen the potential pool of users (thus, contributors) to CorCenCC in the future. We envisage that the app can play its part in showing that a shift in corpus creation methods is possible, highlighting the important role that crowdsourced contributions can play in constructing user-driven resources and representing diverse groups of language users, in the Welsh context and beyond.
References


Keeping the English dative alternation in the family: a quantitative corpus-based study of spoken data
Gard B Jenset (independent researcher), Barbara McGillivray (Alan Turing Institute / University of Cambridge, UK) and Michael Rundell (Lexicography MasterClass / Macmillan Dictionary, UK)

Background and motivation

This talk presents a quantitative study on the English dative alternation construction based on the Early-Access Subset of the Spoken British National Corpus 2014. The features of the corpus data and metadata have allowed us to investigate the grammatical and sociolinguistic factors affecting the speakers’ choice between the V-NP-NP construction (Give me the money) and the V-NP-PP(to) construction (Give the money to me).

Previous research on the dative alternation has focussed on introspection, written texts, or telephone conversations between strangers. This leaves open the question of the interplay between grammatical and sociolinguistic variables in spontaneous speech. Arnold et al. (2000) showed that the structural complexity of constituents and information status play a role in this alternation. Bresnan et al. (2007) found a greater range of possible variation using multivariate statistical techniques, and confirmed the role played by discourse factors. Jenset and Johansson (2013), using data from the web, demonstrated the effects of the semantics of the theme, but the data for this study were restricted to sentences with pronominal recipients. On the other hand, previous sociolinguistic studies on written data (Kendall, Bresnan, and Van Herk, 2011) and spoken data from telephone conversations between strangers (Bresnan and Hay, 2008; and Bresnan and Ford, 2010) found no significant role played by gender and age in the dative alternation. In this study we investigated whether in-person interactions between friends and relatives would lead to different results.

Data

We focussed on the relevant concordance lines of six high-frequency verbs (give, lend, offer, sell, send, and show), which correspond to 1,938 observations in the 2014 Spoken BNC data, containing transcripts of spontaneous conversations in informal settings recorded between 2012 and 2015. Since our attempts at automatically parsing the data yielded poor results due to the low performance of available syntactic parsers on spoken data, we manually annotated the type of dative pattern found in each, whenever relevant, and annotated the head of the noun phrases for recipients and themes.

Further, we automatically collected relevant corpus data and metadata, including speakers’ information (gender, social status, age, etc.) and the semantic class of the heads of the recipient and theme.
Results

The table below displays the relative frequency of the two patterns for each of the six verbs analysed, showing that the V-NP-NP pattern is dominant for all the verbs except sell.

<table>
<thead>
<tr>
<th>Verb</th>
<th>Datives</th>
<th>V-NP-NP</th>
<th>V-NP-PP</th>
</tr>
</thead>
<tbody>
<tr>
<td>give</td>
<td>1000</td>
<td>882 (88%)</td>
<td>118 (12%)</td>
</tr>
<tr>
<td>lend</td>
<td>38</td>
<td>27 (71%)</td>
<td>11 (29%)</td>
</tr>
<tr>
<td>offer</td>
<td>72</td>
<td>62 (86%)</td>
<td>10 (14%)</td>
</tr>
<tr>
<td>sell</td>
<td>103</td>
<td>26 (25%)</td>
<td>77 (75%)</td>
</tr>
<tr>
<td>send</td>
<td>570</td>
<td>436 (76%)</td>
<td>134 (24%)</td>
</tr>
<tr>
<td>show</td>
<td>276</td>
<td>242 (88%)</td>
<td>34 (12%)</td>
</tr>
</tbody>
</table>

To assess the effects of linguistic and sociolinguistic features on the choice of one pattern over the other, we carried out a multivariate analysis. We fitted a binary logistic mixed-effects model with the response being realization of the recipient as either a PP (1/True) or NP (0/False), a random effect for verbs, and the following predictors: gender, pronominality of recipient, pronominality of theme, logarithm of recipient’s length (in number of characters), logarithm of theme’s length, animacy of recipient. By inspecting the predictor’s coefficient, we found that all predictors were statistically significant. Specifically, in our data men were 14% more likely to use a V-NP-PP construction than women. Pronoun recipients were associated with a 50% lower probability of finding V-NP-PP constructions, while pronoun themes were associated with a 50% increase in the use of V-NP-PP constructions. A one-unit increase in recipient length (on a log scale) results in a 35% increase in the use of V-NP-PP, while a one-unit increase in theme length (on a log scale) results in a 50% decrease in use of V-NP-PP. Compared to inanimate recipients, animate ones have a 17% higher rate of use of V-NP-PP constructions. Moreover, considering verbs (the random effects), show and give prefer NP recipients, whereas sell prefers PP recipients.

In line with previous research (Bresnan and Hay 2008; Bresnan and Ford 2010; Kendall et al. 2011), our results show that sociolinguistic variables have a weaker effect on the dative alternation compared to the grammatical ones.

Compared to Bresnan et al. (2007), who focussed on American English, our model downplayed the effect of recipient’s animacy, and had no place for the indefiniteness of theme (omitted for reasons of parsimony). Consistent with their results, we found strong effects for the length of theme and argument and the pronominality of recipient. Moreover, we found a strong effect for the theme’s pronominality; this points to Behagel’s Law and is in line with previous studies on the role played by discourse features on alternating constructions (Arnold et al. 2000; Bresnan et al. 2007; Bresnan and Hay 2008; Bresnan and Ford 2010; Jenset and Johansson 2013).

We found that even in spontaneous conversations between familiar speakers, the main results from previous research held up. However, we found a gender effect that runs parallel to the grammatical effects: while the V-NP-NP pattern is associated with
pronoun recipients and female speakers, the V-NP-PP pattern is associated with non-pronoun recipients and male speakers. This suggests that the gender effect may not be a direct result of specific preferences for one construction over another, but instead an emergent effect arising as an indirect consequence of other linguistic choices.

References


Project Background

The BT Archives house the records of British Telecom, the world’s oldest telecommunications company, which traces its history back to the formation of the Electric Telegraphy Company in 1846. Prior to its privatisation in 1984, BT was a public corporation (and before that a government department) and as a result all of the pre-privatisation material in the archives is in the public domain, making it ideal for academic research. Despite this legal availability, however, the physical availability of material in the archive was limited to two days a week in an archive space in Holborn, London. In 2011 the ‘New Connections’ project was set up with the aim of making around half a million items from the public archives of British Telecom available in a new digital archive. As part of ‘New Connections’, three academic research projects were funded, one of which was the creation and analysis of the British Telecom Correspondence Corpus (BTCC).

The era that the archive covers makes it a potentially fascinating source of data for the linguistic study of business correspondence. The mid-nineteenth to late-twentieth century is a crucial period in the development of English business correspondence as the amount of business being conducted by letter increased massively during this period as a result of the Industrial Revolution, the introduction of the Penny Post, and increased access to education both in schools and through composition grammar guides. Despite its importance in the development of business correspondence, this period has received relatively little attention. The aim of constructing the British Telecom Correspondence Corpus was to start addressing this gap in available linguistic data and enable studies into the development of business correspondence from the mid-nineteenth to late-twentieth century.

Creating the Corpus

The first major challenge in creating the corpus was to identify letters. The material for the digital archive had been preselected by British Telecom, and was digitised and delivered as image scans to Coventry University by The National Archives. In this sense, the initial data collection had already been done. However, the digitised files were labelled and organised according to British Telecom and Post Office archive finding numbers (e.g. TCB 473/P 10045) and contained little or no metadata. The only way to identify letters was, firstly, to set criteria as to what constituted a ‘letter’ for our purposes, then search through around 13,000 individual scans manually. This first phase of searching identified just over 500 letters. Once the digital BT Digital Archive launched it was possible to search for additional letters but again the lack of item-level metadata meant that similar challenges persisted. Overall 612 letters were selected for the corpus, which in its current state contains just over 130,000 words.
The letters are distributed relatively evenly across the fourteen decades represented. An attempt was made to include a wide variety of authors, occupations, and companies as possible to get as wide a representation of the available material as possible. The correspondence is mostly written in British English, though as British Telecom and the Post Office did a great deal of international business there are letters from many different locations, perhaps most notably America from which there are letters regarding, for example, the first trans-Atlantic Telephone calls and co-operation over satellite testing. The authors are largely male despite an initial hope that the corpus might be more balanced in terms of gender. This imbalance seems largely due to the roles that women typically held in telecommunications companies during this period which did not generally involve authoring correspondence. The data was transcribed through a mixture of manual transcription and OCR (Optical Character Recognition) scans of type-written material.

Finally, the letters were classified in relation to the overall pragmatic function they serve. A relatively small list of ten functions (such as Applications, Requests, Commissives, Offers...) was used to try and limit the effects of data scarcity, while offering a starting point for comparison of the various sorts of letters represented in the corpus. The classification also offered an additional way of interpreting the purely quantitative results (for example the n-gram ‘let me know’ appears most frequently in Queries).

Methods of Analysis

The analysis of the corpus was primarily data-driven, that is to as few preconceptions as possible were made about the data, and I worked from the starting point that, as Stubbs put it, ‘repeated events are significant’ (2007: 130). This approach seemed consistent with the core principles of corpus linguistics, and appropriate for an essentially exploratory study. I generated a list of n-grams of between three and six words that appeared more than 25 times in the corpus as a whole, and examined each for stability and change with regard to frequency and function over the timeline of the corpus. In addition to this keywords were generated using each decade as a sub-corpus and comparing against the whole corpus. These keywords were organised according to the categories outlined by Scott (2012): proper names, indicators of ‘aboutness’, and potential indicators of style.

The majority of the keywords (66%) were indicators of ‘aboutness’. Many of these results were interesting in terms of topic but of limited linguistic interest and so were used to help contextualise findings in terms of topic. Each of the potential style-marker keywords, which made up 12% of the keywords overall, were examined in context for how they were used and whether the patterns in which they appeared changed over time.

Some Initial Findings: Corporate Identity

Decline in Deference: Formalised distance to formalised friendliness

One clear pattern highlighted by the quantitative results is a decline in overtly polite and deferential terms of address. Seven of the most frequent n-grams are variations
on the formulaic closing ‘your obedient servant’, which is most often preceded by another frequent three-word n-gram ‘I am sir’, the seventh most frequent three-to-six-word n-gram in the BTCC. Overall these formulas decline in frequency in the early-mid twentieth century and disappear from the corpus in the late 1950s.

We also see a corresponding decline in the use of formal opening terms of address such as ‘Sir’ in favour of the use of named recipients (e.g. ‘Dear Ken’). In the latter half of the twentieth century ‘Dear [first name]’ is increasingly the most popular opening formula, and ‘yours sincerely’ dominates closing formulas. Similarly, where authors in the late nineteenth and early twentieth century tended to manage the exchange of letters with phrases such as ‘with reference to your letter’, increasingly authors write ‘thank you for your letter’. This seems to be part of a wider decline in more negatively polite (Brown and Levinson, 1987) forms in favour of positively polite forms of address, which historically have been more typical of personal correspondence. While these forms are familiar on the surface, however, they also occur in increasingly standardised forms with no variation to indicate degrees of social proximity. Overall this suggests a move from formalised distance towards a sort of formalised friendliness.

Decline of pre-performatives and the emergence of the secretarial ‘we’

Two further trends are identified by the quantitative analyses. First of all there is something of a decline in the first person pronoun ‘I’ and an overall increase in the first person plural pronoun ‘we’. When examined in closer detail, this appears to be related to the decline in pre-performative phrases such as ‘I am directed to’, ‘I am to’, and ‘I beg to’. Where previously secretarial authors performed their role as conveyor of the message, as in example (1),

(1) "I am directed by the Postmaster General to acknowledge the receipt of your letter of the 7th ultimo...” (1870_04_02_FIS_DHC)

Increasingly authors made the distinction between themselves as the conveyor of the message and the corporate body on whose behalf they write with personal pronouns (i.e. ‘I’ as conveyor of the message, ‘we’ as source of the message).

This decline in pre-performative clusters occurs in the BTCC around the same time as the decline in deferential formulas such as ‘your obedient servant’. The overall effect of these changes is that we end up with a business language that is more democratised in the sense that the corporate and personal hierarchies are not explicitly performed. However, this democratisation also seems to have contributed to the impression that business language has become more impersonal, and that there has been a shift ‘from the individual to the corporate dimension of letter-writing’ (Del Lungo Camiciotti, 2006:171).

Research on the corpus is ongoing, with the inclusion of additional material planned for late 2017. For now the BTCC has started to fill the gap in historical business correspondence data for this period, and provided a new way of engaging with the rich historical material in the BT Archive.
References


As a method of linguistic inquiry, corpus linguistics relies on corpora as a sample of a larger population. As with all other scientific disciplines, empirical findings from a sample can only be generalized to a larger population if the sample is representative of that population. According to Biber (1993a), the representativeness of a corpus is determined by “the extent to which a sample includes the full range of variability in a population” (p. 244). Representativeness in corpus design is crucial since the goal of most corpus studies is to identify quantitative linguistic patterns in the corpus sample and generalize those findings to a larger linguistic population.

A fuller understanding of how issues of corpus design impact its ability to represent a larger population is particularly important as we witness marked increases in the publication of empirical linguistic research based on corpora (Sampson, 2013). Yet only a few studies (Biber, 1990, 1993a, 1993b; Gries, 2008) have addressed this issue empirically; most explicit treatments of representativeness instead focus on general recommendations (e.g., Váradi, 2001; Gries, 2006; Leech, 2007), and some even advocate that sample size is the most important aspect of corpus design (Sinclair, 1991; Hanks, 2012).

From a statistical sampling perspective, this overemphasis on size at the expense of other sampling considerations raises serious questions about the validity of many corpora and the published findings based on them. Unfortunately, to date, there have been no large scale studies that have empirically tested the impact of sampling decisions and corpus size on the representativeness of corpora, nor have there been empirical evaluations of representativeness of existing corpora. In this presentation, we begin to fill this gap with the results of two case studies empirically evaluating corpus representativeness.

Conceptually and methodologically, there are two major types of representativeness: target domain and linguistic representativeness (Biber, 1993; McEnery, Xiao & Tono, 2006). We define target domain representativeness as the extent to which a corpus contains the full range of text type variability that exists in the target domain. Target domain representativeness determines the generalizability of a corpus sample to a larger population of interest. We define linguistic representativeness as the extent to which a corpus contains the full range of linguistic distributions that exist in the target domain. Linguistic representativeness determines the suitability of a corpus sample for answering specific research questions about specific linguistic features. Importantly, linguistic representativeness is inherently related to the linguistic level being investigated; the same corpus may be representative of a common grammatical structure, but not of lexical distributions.

---

1 Target domain representativeness has also been referred to as external (e.g., McEnery, Xiao & Tono, 2006) or situational (e.g., Biber, 1993) representativeness. Linguistic representativeness has also been referred to as internal representativeness (e.g., McEnery, Xiao & Tono, 2006)
A major methodological challenge for evaluating corpus representativeness is estimating the linguistic and situational characteristics of the target population. After all, corpora are created in part because of the inability to study all of the language of a particular type. There are a few cases in which we can actually study the full population in addition to smaller samples of that population. For example, it is possible to download the entire body of Wikipedia articles (c. 5.3 million articles; 2.4 billion words). On the other hand, it would not be feasible to collect every research article published in a discipline. Thus, different methods are required to empirically evaluate representativeness when the full population is a known entity versus when it is not (which is the case for most corpora in use today).

Table 1 summarizes the two parameters that must be taken into account in evaluating corpus representativeness (1 and 2) and whether or not the full population can be analyzed (A and B). The table also proposes one possible method through which an evaluation of representativeness could be approached (other methods are also possible) in each scenario.

Table 1. Possible methods for evaluating corpus representativeness

<table>
<thead>
<tr>
<th></th>
<th>A. Full Target Population Unknown</th>
<th>B. Full Target Population Known</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Target Domain Representativeness</td>
<td>A1. Carry out a detailed analysis of the situational characteristics of each text in a corpus. Compare these characteristics to a survey of the target domain, or extrapolate what population the corpus can be generalized to.</td>
<td>B1. Compare the occurrence and/or proportion of situational characteristics represented in the full population to a series of experimental corpora that represent different methods of corpus construction.</td>
</tr>
<tr>
<td>2. Linguistic Representativeness</td>
<td>A2. Divide an existing corpus into smaller, random samples. Compare the dispersions of linguistic features across the smaller samples and to the full corpus.</td>
<td>B2. Compare the distributions of a range of linguistic features in a corpus containing the full population to a series of experimental corpora that represent different methods of corpus construction.</td>
</tr>
</tbody>
</table>

In an ongoing project (Egbert, Gray, & Biber, under contract), we carry out evaluations of representativeness in each of these categories. In this presentation, we use the results of two of these as case studies.

**Case Study 1. Evaluating Target Domain Representativeness in a Corpus of Academic Research Articles**

In the first case study, the issue of target domain representativeness is addressed when it is not possible to analyze the full target population. In this case study, target domain representativeness is assessed for a 270-text corpus of research articles (c. 2 million words) in 6 disciplines: philosophy, history, political science, applied linguistics, biology, and physics. This case study represents a cyclical process including target domain analysis, corpus design and compilation, and corpus documentation and evaluation (via situational analysis of the corpus texts). It is the final step, a comprehensive analysis of the situational characteristics of the texts included in the corpus, that enables an evaluation of the target domain representativeness of the corpus. In the first stage, a comprehensive survey of
academic journal registers is carried out to identify text types across journals and registers, and to identify operational definitions for identifying those text types. In the second stage, the results of the target domain survey are used to identify the sampling categories for the corpus, and used to collect a stratified corpus that is balanced across sub-corpora. In the final step, all texts in the corpus are analyzed for their situational characteristics (e.g., discipline, type of research, organization pattern, explicitness of research design, nature of data, topic, etc.), and compared to the results of the target domain survey.

**Case Study 2: Evaluating Linguistic Representativeness in Corpora of Wikipedia Articles**

This case study compares the full population of Wikipedia articles (c. 5.3 million articles; 2.4 billion words) to a series of sixteen experimental corpora created from the full Wikipedia population. The experimental corpora are built using four different sampling methods and four size thresholds. Two probability sampling methods (simple; stratified) are included, along with two non-probability sampling methods (convenience; quota). For each of the four sampling methods, we have sampled four corpora of differing \( N \) sizes, measured in the percent of the Wikipedia population that is sampled (.001%; .01%; .1%; 1%). These four sample sizes range from very small samples of about 53 texts (1/100,000 of the population) to large samples of about 53,000 texts (1/100 of the population). The composition of the experimental corpora is documented in Table 2.

<table>
<thead>
<tr>
<th>Type</th>
<th>Probability</th>
<th>Non-probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>Simple</td>
<td>Stratified</td>
</tr>
<tr>
<td>Size</td>
<td>.001% (c. 53 texts)</td>
<td>.001% (c. 53 texts)</td>
</tr>
<tr>
<td>((N))</td>
<td>.01% (c. 530 texts)</td>
<td>.01% (c. 530 texts)</td>
</tr>
<tr>
<td></td>
<td>.1% (c. 5,300 texts)</td>
<td>.1% (c. 5,300 texts)</td>
</tr>
<tr>
<td></td>
<td>1% (c. 53,000 texts)</td>
<td>1% (c. 53,000 texts)</td>
</tr>
</tbody>
</table>

The full Wikipedia corpus and each experimental corpus is analyzed for a range of linguistic variables, including frequency counts for lists of words in three categories: high-frequency function words (e.g. the, an, of, to), mid-frequency content words (e.g. home, computer, break, teach), and low-frequency technical words (e.g. metastasize, sedimentation, semiconductor, altruistic), and grammatical features in two categories: high-frequency part of speech classes (e.g. nouns, verbs, adjectives, adverbs) and low-frequency grammatical structures (e.g. that-complementizers, passive voice). Each of these features is measured in terms of normed rates of occurrence (per 1,000 words) in each text. Means and standard deviations across texts in each sample are used to evaluate the extent of linguistic representativeness of the experimental corpora compared to the full Wikipedia population.

The results of the case studies are discussed in terms of their implications for corpus design in other domains of language use. We argue that evaluations of corpus representativeness should become more explicit and transparent, and briefly discuss additional methods and the challenges of carrying out such evaluations. In addition, recommendations are given for documenting and disseminating information about corpus design and representativeness.
References


In this work we present a probabilistic method to quantify the distance of a corpus from the standard model. We then apply this metric to show that the degree to which the actual language use deviates from the standard language varies significantly from region to region. Specifically, with the Twitter data coming from four broad regions in Japan, Tohoku, Kansai, Chugoku and Kyushu (about 17,000 tweets each), we observe widely differing distances, which indicates that the readiness to conform to the standard language varies between regions.

We started by training a model with a newspaper corpus written in standard Japanese and a lexicon of standard Japanese. For this training, as well as for computing the plausibility of a sentence, we use the morphological analysis tool MeCab (Kudo, Yamamoto, & Matsumoto, 2004). This tool employs the Conditional Random Fields (CRF) algorithm (Lafferty, McCallum, & Pereira, 2001) to build the feature-based statistical model, and computes the ‘cost’ based on the probability of a sentence (the smaller the cost the greater the conformity to the model). CRF, in its training, takes into account not just the words’ surface forms but their various attributes, as well as their order. We used in our training nine features including lemma and part of speech. MeCab further considers multiple paths for different word segmentation hypotheses. This functionality of MeCab, designed primarily to adapt to the texts in an orthography with no space between words, nicely mimics the real situation in speech where no apparent word segmentation is given. Thus, we are closely modelling the situation where a nono-dialectal speaker is exposed to a non-standard dialectal sentence, which may pose him/her a variety of challenges related to either grammar, vocabulary or word segmentation.

We then confirmed the expectation that on average, greater costs are in fact associated with dialects, using pre-classified parallel dialect corpora developed by (Yoshino et al., 2016). The corpora consist of a set of standard Japanese sentences as well as its translations into the dialects of the above four regions, all hand-crafted to ensure that there will be great contrasts between dialects. Reflecting this intended contrast, the average cost for the standard-Japanese is −1, 570, while that for each dialect is much higher (over 30,000).

Taking these numbers as expected costs, we then processed our region-classified Twitter data and compared their costs. The distance for each region is calculated as the expected cost minus the twitter cost, producing the following figures:

Tohoku 56, 148 − 30, 301 = 25, 847
Kansai 31, 646 − 32, 279 = −633
Thus the results show that for our four regions, the size of deviation from the standard language varies from one another. In particular, Kansai dialect speakers seem the most willing to adhere to their own dialect, while Tohoku dialect speakers are the least willing. This confirms the conventional observation by the Japanese people.

We also compared the differences in cost of four regions inside the twitter corpora against the standard Japanese speaking region (Kanto, at 29,747). The ANOVA analysis shows that the regional deviations from this figure are mostly significant (P < .01) with the exception of Tohoku dialect speakers, reinforcing the aforementioned contrast in regional characteristics in terms of willingness/reluctance to speak their dialect.

Our method of measuring distances between subgroups of a corpus is a general one, in the sense that it can be straightforwardly extended to other languages to test their diversity, as long as there is a 'standard' set of resources (lexicon and annotated corpus). Also, the target parameter does not have to be dialect, and could be of any genre: one could for example apply the method to see if the language use of a social subgroup diverges from the society as a whole.

The main limitation of our method on the other hand is that it only shows the divergence from a certain standard, not the differences between any given set of groups. Also, there is no guarantee either that the distance we observed is actually due to a dialectal difference (if likely). To be able to distinguish between any regions, i.e. cluster dialects, we plan therefore to complement it with other metrics, such as vocabulary share rate (as in (Inoue, 2008)).

References


Aims and background

Frequency plays a central role in corpus and usage-based linguistics. However, there remain many open questions about the role of diverse frequency types in language learning, processing and production (cf. Divjak & Gries 2012). The aim of the present paper is to contribute to a more exact understanding of the role of different frequency measures in shaping up language structure. In particular, we focus on the well-known correlation between word frequency and length. This correlation has been well-known since Zipf’s seminal work (1935[1968]) as the Law of Abbreviation. It has been shown that the Law of Abbreviation is an absolute language universal (Bentz & Ferrer-i-Cancho 2016).

More recently, a study by Piantadosi et al. (2011) has demonstrated that the average contextual informativity of a word (i.e. the inverse of the conditional probability of the word given the preceding n-grams) in fact correlates more strongly with the word length than the context-free probability (i.e. normalized token frequency). More predictable (and therefore less informative) words tend to be shorter, whereas less predictable (and more informative) words are usually longer. These important findings call for a new evaluation of Zipf’s legacy and support the theory of uniform informational density as a means of optimization of human communication (e.g. Jaeger 2010).

At the same time, it remains unclear how these findings tie in with a well-known phenomenon in typology, namely, formal asymmetries between members of grammatical categories (e.g. Greenberg 1966). For example, it is well known that singular nouns tend to be formally unmarked or less marked than the corresponding plural forms across a variety of languages (e.g. Brunner 2010), e.g. English chair – chairs, German Stuhl – Stühle, Russian stul-Ø “chair.NOM” – stul-ja “chairs.NOM”. This fact has been explained by frequency asymmetries between the forms, i.e. by the tendency of the less formally marked forms to have higher relative frequencies than the more marked ones (Haspelmath 2008), in accordance with the principle of economy and minimization of effort. The goal of the present study is to find out whether context-based informativity is more strongly associated with grammatical asymmetries than context-free relative frequency. For this purpose, we perform three case studies, focusing on 1) singular and plural nouns, as in the examples above, 2) positive, comparative and superlative forms of adjectives, e.g. fine – finer – finest, and 3) cardinal and ordinal numerals, e.g. ten – tenth.

Data and method

In these case studies, we employ the Google Books n-grams in English, French, German, Italian, Russian and Spanish (Lin et al. 2012), which are available online at http://storage.googleapis.com/books/ngrams/books/datasetsv2.html (last access 09.01.2017). We select samples of nouns, adjectives and numerals (100–400 words
per class) and compute the relative (paradigmatically) frequencies, as well as average contextual informativity scores, for the singular and plural forms (nouns), different degrees of comparisons (adjectives), and ordinal and cardinal forms (numerals). Lexemes with identical forms, e.g. German Lehrer “teacher (SG and PL)”, are omitted because of the lack of sufficient morphological information in the data. Following the results presented in Piatandosi et al. (2011), the informativity scores are based on the left-context $n$-grams with $n=3$. To determine which of the two measures is more strongly associated with the formal asymmetries between the less and more marked categories, we use non-parametric paired Wilcoxon tests and mixed-effects binomial and ordinal logistic models. In these models, the contrasting category members (e.g. singular or plural) serve as the response, and the frequency and informativity measures function as predictors. The lexemes (specific nouns, adjectives and numerals) are treated as random effects (intercepts).

**Preliminary results**

Our preliminary results show that both frequency and informativity behave as expected across the languages: the unmarked and shorter forms are usually more frequent and less informative than the marked and longer ones. Figures 1 and 2, which represent the results for the singular and plural forms of nouns in British English, illustrate the point. Moreover, frequency and informativity are strongly correlated. However, unlike in the study of lexical units in Piatandosi et al. (2012), the more sophisticated informativity measures do not yield a significant improvement in the discrimination between the shorter and longer grammatical categories in comparison with the simpler frequency measures, sometimes performing even worse, as in the illustration. The results thus suggest that informational density is not the only factor that determines the linguistic form and that one should take into account the paradigmatic relationships between forms and categories.

![Figure 1](image_url)

**Figure 1.** The difference in log frequency between singular and plural nouns in British English. Wilcoxon paired signed rank test: $V = 15436, p < 0.001$. 
Figure 2. The difference in informativity between singular and plural nouns in British English, based on 3-grams (left context). Wilcoxon paired signed rank test: $V = 8644, p < 0.014$.

References


The meaning-making processes in human interactions are inherently multimodal and are usually mediated by more than one mode of communication (Adolphs & Carter, 2013; Knight, 2011; Seedhouse & Knight, 2016). Whereas enquiries of the speech mode, usually based on textual transcripts of utterances with a varying degree of granularity, are already a well-established strand of research in Corpus Linguistics (CL), multimodal corpus-based examination and understanding of other modes such as prosodic features and bodily movements (e.g. postures, facial expressions, head and hand-and-arm movements etc.) still remain underexplored.

Despite the affordances provided by tools and software for textual annotation, exploration and description, these methods are mostly unsuited for the detection and description of non-verbal behaviours. While multimodal alignment and annotation tools such as ELAN (Wittenburg, Brugman, Russel, Klassmann, & Sloetjes, 2006) have proven to be invaluable to Gesture Studies, (semi-)automatic and fine-grained gesture demarcation and pattern detection technologies are still largely unavailable. To enhance the efficiency and robustness of multimodal linguistic research, the issue of how to integrate existing mainstream corpus techniques into multimodal analysis of human communication becomes vital.

To address the theoretical and methodological issues mentioned above, the current paper will focus on demonstrating approaches and findings of a systematic multimodal corpus-based study of a recurrent group of gestures i.e. circular gestures (CGs) in relation to the speech that co-occurs with them. Expanding the recently proposed concept of ‘cyclic gestures’ (Ladewig, 2014; Ladewig & Bressem, 2013), the Circular Gesture (CG) refers to any continuous hand-and-arm movements that depict circles.

The relationship between CGs and speech is being examined from three aspects, i.e. temporal alignment, speech patterns and cognitive motivations. The research draws on a sub-corpus of the Nottingham Multimodal Corpus (NMMC), a collection of 8 videos (about 8 hours) of real-life postgraduate supervision meetings. In total, 570 instances of CGs have been identified and annotated in Elan. To enable and facilitate the investigation of speech patterns, all of the speech components aligned with the CGs were annotated with “CG# #”, with the speech located in-between the hashtags. Using CG as a search item, all of the textual instances are then ready to be automatically searched, sorted and analysed in any corpus tool. We will demonstrate this using AntConc (http://www.laurenceanthony.net/).

Based on the Concordance list generated in AntConc, the temporal analysis between the CGs and the speech reveals that the CGs tend co-occur with multiple lexical items and distinct patterning. Searching for the most frequently occurring clusters to the left of the node “CG” for example
foregrounds a large number of instances of speech with short pauses in the middle of dysfluent utterances.

With the assistance of the alphabetically formatted concordance of the speech parts that co-occur with the CGs, the semantic and grammatical variations can be described and analysed qualitatively and quantitatively. The current paper will concentrate on the category with the largest number of instances, all of which seem to be associated with the contexts where the meaning of continuity, progress and progression of time, events, actions and thinking is indicated. This finding is consistent with research on German (Ladewig, 2011, 2014) and French corpora (Calbris, 2011), which points to the universal meaning of CGs across languages.

In terms of the cognitive motivation underlying such a gesture-speech association, we can refer to the metaphoric image schema that **MIND IS A MACHINE, THINKING IS A PROCESS IN A MACHINE, AND TIME IS MOTION THROUGH SPACE** (Ladewig, 2011; Lakoff & Johnson, 2008). The cognitive motivation embedded in this image schema is most likely entrenched throughout repetitious real-life experiences of the physical body in prevailing phenomena such as the continuous rotation of the wheels and machines, and the constant rising and falling of the sun, etc.

Whereas the previous studies have limited their description of the functions of CGs to semantics (e.g. expressing the meaning of on-going process), cognition (e.g. retrieving words during dysfluent speech) and pragmatics (e.g. encouraging the interlocutor to take the floor), the current paper proposes a description that focuses on CGs’ multiple discourse functions. These include contributions to meaning-making, segmenting utterances, emphasising noteworthy information, maintaining coherence and cohesion and enhancing inter-subjectivity.

The paper demonstrates how the combined application of multimodal annotation tools and existing corpus software can significantly enhance the efficiency and reliability of multimodal investigations of emerging patterns of gesture-speech relationships. In addition, we show how corpus technologies can be further developed in order to contribute more fully to the examination and understanding of multimodal interactions in a wide range of communicative contexts.

**References**


Investigating L2 errors in a quasi-longitudinal learner English corpus, with particular reference to word order and the position of also
Amanda Murphy and Francesca Poli (Università Cattolica del Sacro Cuore, Italy)

The motivation for this study comes from university language teachers’ observations of Italian EFL students’ ‘non-target use’ of certain elements or structures (i.e. errors) in written English texts. On the basis of the experience of correcting scripts by EFL students, teachers identified certain errors as persistent, particularly, incorrect choice of prepositions, usage of the present perfect, past simple tense, conditionals, and word order. The intuition of the presence of the Second Language Acquisition concept of fossilisation, understood by Han (2004: 13) both as a process (the premature cessation of development in defiance of optimal learning conditions) and as a product (structures that persist over time, against any environmental influences) began to trigger some research questions. The teachers began to wonder when these errors arise, and devised a project on learner language as it develops through time.

In an era when the debate around English as a lingua franca continues to grow (Di Scala, 2016), and new varieties of English are regularly coded and described by international scholars (starting from the ICE project – Greenbaum, 1991), it may seem curious to insist on correct or incorrect usage of English. However, as Granger robustly argues (2015: 495), it is early days to specify new norms in English, and in a school/university learning context it is not unreasonable to investigate and point out incorrect usage, even more so in a faculty of foreign languages where teaching staff and students alike would like to strive for excellence.

The first broad research question asks whether the five errors - incorrect choice of prepositions, usage of the present perfect, past simple tense, conditionals, and word order - occur continuously throughout written learner production from school to university, or whether they diminish in frequency. The second, stemming from the first, investigates the specific case of the additive adverb also, seen in Italian native speaker written learner production to be often misplaced in the clause.

To trace the history of such errors, it was decided to compile what Granger (2008: 262) calls a pseudo or quasi-longitudinal corpus, which collects data from different learners with different levels. A learner corpus of written EFL texts was thus collected between February 2013 and October 2014 and divided into four sub-corpora, corresponding to texts from middle school, secondary school, undergraduate and postgraduate university students. Texts in the Undergraduate Corpus were written in the first of the three BA years, and the texts in the Postgraduate Corpus were written in the first year of a two-year MA degree.

No judgement was made as to the students’ level of proficiency; the texts were grouped by external criteria. The corpus was compiled ‘as it was possible’, i.e. local teachers from the different school levels were contacted, and asked for examples of student texts to include in the corpus. The only requirement given, following Granger (2002: 8), was that “the language sample must consist of continuous stretches of discourse, not isolated sentences or words”. No short exercises or translations were accepted. As most of the texts were handwritten, they were digitalised and coded according to school level (Middle School, Secondary School, etc.). The Middle School sub-corporus comprises 35 texts for a total of 31,480 words, the Secondary School sub-corporus 258 texts for a total of 53,678 words, the University sub-corpora consist of 47 texts for a total of 13,939 words; subdivided into 33 Undergraduate texts totalling 3,675 words and 14 Postgraduate texts, totalling 10,264 words. The resulting corpus is clearly not balanced, but nevertheless provides data that can be used for broad comparative research.

To cross-check whether the five errors to be investigated were indeed relevant, a sample of the corpus (10% of the 340 texts in the corpus, which amounts to 100,000) was
submitted to four other colleagues, two non-native speaker teachers of English in school and two native speaker teachers of English at university) with the request to categorise errors in the corpus.

The raters identified two key patterns, one positive and one negative: the former demonstrates that peak incorrect usage of the two past tenses and of the conditionals occurs at high school and middle school level respectively, followed by gradual improvement in the more advanced stages of learning. The negative pattern identified that incorrect usage of the prepositions and word order errors peaked at high school and postgraduate level. Thus, these findings offered the grounds for a more in-depth analysis of the corpus in order to verify whether there is indeed improvement in the use of tenses and conditionals, and a decline in the use of prepositions and word order.

Each text in the corpus was then manually read for the five errors under investigation, firstly by the principal researcher, then by one of two raters, both non-native teachers of English, who categorised errors. A score was assigned to each text for the total number and type of errors, and these scores were subsequently compared across the sub-corpora. Where their scores differed, an average was made of the errors they identified.

The error proportion was calculated with two variables acting as the main operands in the formula: the number of errors (E) and the total number of sentences (S):

\[
\frac{E}{S} \times 100
\]

Thus, the number of errors (e.g. 1) was divided by the total number of sentences (e.g. 2), obtaining the values of 0.5, which was subsequently multiplied by 100 resulting in the final error proportion of 50%. This simple formula allows for easy and immediate comprehension of the data. The error proportion was calculated for the texts and the percentages were added up and divided by the total number of texts of that specific sub-corpus, thus obtaining an average error proportion for each educational level. The single values for each level for each of the structures analysed in the research were then used to carry out a comparison (in the form of a graph) between the texts.

A comparison of the results revealed that except for word order, the proportionate frequency of the other four categories of errors diminished as the level of education rose, i.e. middle school students incorrectly used verbs in the past simple, present perfect, and the conditional, and prepositions more than postgraduate students. The only category which did not improve was that of word order, as Graph 1 illustrates.
Overall, in the Middle School Corpus, 18/35 texts (51.42%) contained word order mistakes; in the Secondary School Corpus 112/231 texts (48.18%); in the Undergraduate Corpus, 18 out of 33 texts (54.54%), and all 12 Postgraduate texts (100%).

The hypothesised reason for this finding was that the Postgraduate texts under analysis were more complex, with long sentences. This hypothesis finds some evidence in the texts, since the Postgraduate texts were short academic essays, requiring some argumentation. Perhaps Selinker’s 1972 hypothesis that as students concentrate on content, their attention to form diminishes is also correct, although this cannot be proved.

The second investigation into the learner language in the corpus regarded the use of the additive adverb also, which was noted in the reading of the texts as often being misplaced in the clause, especially in the Postgraduate Corpus.

This may not be surprising, since reference grammars such as the Longman Grammar of Spoken and Written English warns readers that the meaning of additive adverbs can change, according to its position in the clause (Biber et al. 1999: 781)

Investigations of the adverb also using the concordance software AntConc (Anthony, 2011) produced examples that showed the common tendency for also to be placed directly after the verb in the Postgraduate Corpus, as in:

(1) ‘the Fall of the Berlin Wall had also (WO) negative consequences’;
(2) ‘A widespread system of communication improves also (WO) our possibilities to know the world’

Quasi-longitudinal learner corpora can provide a bird’s-eye view of the overall broad progression of learner language, although detailed comparisons between levels need to be made with caution. Obviously, the larger the corpus, the more representative the results would be, and the unbalanced nature of the present corpus is a limitation of the present research. Nevertheless, the investigation opens up interesting questions as to possible areas of fossilisation among learners, particularly at a higher level; according to Selinker’s early study (1972: 215), fossilised structures tend to remain as potential performance, reappearing in learners’ production even when thought to have been eradicated. Findings such as those from the present study could direct teachers of English to areas of language use, such as word
order, and the placing of particular adverbs, such as also, whose difficulty is perhaps under-recognised and deserves attention. A hypothesis for the frequent incorrect use of also could be negative transfer (Selinker and Gass, 2008) from the students’ L1, in this case Italian, but that would require research into contrastive analysis hypotheses and learner corpora, and another paper.

References


**English taste and smell adjectives as generalized emotional vocabulary**

Bodo Winter (University of Birmingham, UK)

**Introduction**

One of the primary functions of language is the expression of evaluation (e.g., Dam-Jensen & Zethsen, 2007; Hunston, 2011). Another important function is the expression of sensory perceptions (e.g., Miller & Johnson-Laird, 1976). These two domains—perception and evaluation—are intimately connected. It has been claimed that compared to relatively neutral visual words such as *yellow* and *purple*, taste and smell adjectives have an obligatory evaluative component (Levinson & Majid, 2014). Whereas visual adjectives may be highly positive or negative (e.g., compare *shiny* and *attractive* to *dark* and *ugly*), there also is a large number of very neutral terms. On the other hand, it is difficult to find taste and smell descriptors that are not highly valenced. For example, the smell descriptors *rancid*, *pungent* and *stinky* have strong negative connotations compared to the highly positive descriptors *fragrant* and *aromatic*.

In this paper, I will combine various perceptual and emotional norm datasets with corpora to look at the semantic prosody of English perceptual words, investigating their evaluative functions in context. Crucially, my aim is to show that taste and smell words form a “generalized emotional vocabulary” that can be used across domains, that is, outside of their primary perceptual contexts, such as the description of food or perfumes. In other words, I propose that the evaluative component of taste and smell adjectives is so strong that they are very frequently used metaphorically to describe non-taste and non-smell domains.

**Analysis of the Corpus of Contemporary American English**

A set of sensory adjectives taken from Lynott and Connell (2009) will be analyzed. To construct this dataset, native speakers of English were asked to norm 400 words according to which perceptual modality they correspond to (e.g., *fragrant* is classified as “smell”, *purple* is classified as “visual”). These modality norms will be combined with emotional valence norms from Warriner, Kuperman and Brysbaert (2013), the SentiWordnet (Esuli & Sebastiani, 2006) and the Twitter Emotion Corpus norms (Mohammad & Kiritchenko, 2015). All of these datasets assign positive and negative numerical values to words according to their emotional valence. Crucially, the valence norms were constructed using different approaches: via human norms (Warriner et al., 2013), via a computational dictionary-based approach (SentiWordNet), and via a corpus-based approach based on occurrences with emotional terms (Mohammad & Kiritchenko, 2015).

In my first analysis, I use these emotional valence measures together with the Corpus of Contemporary American English (COCA, Davies 2008-), a large 520-million-word corpus balanced for five generic registers (newspaper, magazine, academic, literature, spoken). For the set of 400 adjectives, about 150,000 adjective-noun pairs (types) were extracted. A mixed linear regression model with perceptual modality as a fixed effect (sight vs. sound vs. touch vs. taste vs. smell) and word and noun as random effects (with by-modality random slopes) shows that across the entirety of COCA, taste adjectives overall modify more positive nouns and smell adjectives modify more negative nouns. By additionally using an absolute valence measure that quantifies overall emotional involvement (regardless of the sign of the emotional valence) it can furthermore be shown that taste and smell words overall occur with reliably more valenced nouns than words corresponding to
the other sensory modalities. For example, a visual word such as *yellow* frequently modifies relatively neutral nouns such as *trousers* or *house*. These results are obtained regardless for which emotional valence datasets are used.

**Analysis of online review data**

The evaluative meaning of taste and smell words is most apparent for text types in which the main goal of speakers/writers is to express their subjective opinion. For this, online restaurant reviews from the Yelp Dataset Challenge will be used, comprising a set of 2.7 million UK and US restaurant reviews (of both U.K. and U.S. restaurants). Using the Yelp app, users can assign star ratings ("1 star", "2 stars" etc.). Potts (2011) has shown that this allows inferring the emotional meanings of words. Replicating the main findings above, taste words occurred overall in more positive reviews (high star rating) and smell words in more negative reviews (low star rating). Moreover, as the examples in Figure 1 show, review valence (in terms of star rating) yields monotonic increasing or decreasing functions for particular words.

![Normalized frequency of 'rancid', 'reeking', 'sweet', 'tangy' vs Yelp Review Rating](image)

**Figure 1.** Frequencies by rating (5 = most positive), cf. Potts (2011)

So far, the analyses have either aggregated across contexts (COCA) or investigated taste and smell words in food-related contexts (restaurant reviews). To truly show that taste and smell words characterize a domain-general set of affectively loaded terms that can be used across different topics, the current results need to be compared to descriptions for a non-food-related domain. Movie review data is excellent for this because movies are exclusively experienced through vision and sound. As the following sentence from Pang and Lee’s (2004) online movie review dataset shows, movies are often talked about metaphorically in terms of food, using a high number of taste and smell words:

*Watching Trouble Every Day (...) is like biting into what looks like a juicy, delicious plum on a hot summer day and coming away with your mouth full of rotten pulp and living worms*

The perceptual words from Lynott and Connell (2009) were checked against the movie review corpus by Pang and Lee (2004), who categorized 10,000 sentences from the review site www.rottentomatoes.com for whether they were “objective” (occuring in the movie synopsis and describing the movie in a matter-of-fact style) or “subjective” (occurring in the evaluative part of the movie that expresses the critic’s opinion). Analyses of these 10,000 sentences showed that taste and smell words were more likely to occur in the subjective sentences than in the objective sentences. Token frequencies were used to construct odds ratios (subjective versus objective). Taste, smell and touch words were more likely to occur in subjective than objective sentences (3.4, 2.0, 1.5), compared to sight and sound which showed a slight preference for objective sentences (0.7, 0.8). A Poisson regression model shows these differences to be reliable.
Conclusions

Taken together, the findings support Levinson and Majid (2014), as well as Lehrer’s (2009) analysis of perceptual adjectives in wine speech as performing both affective function as well as descriptive function. These results also call into question the proposal that expressions such as sweet melody are “synesthetic” (Shen, 1997; Yu, 2003; Shen & Aisenman, 2008), i.e., mappings between distinct sensory modalities (transfer of taste to sound). Instead of describing a melody as sweet in a truly gustatory sense, such expressions seem to be primarily involving the mapping of affective meanings. Altogether, the results suggest that taste and smell words form a subpart of the English lexicon that characterizes a domain-general emotional vocabulary.

References

Variability in citation practices of developing L2 writers in first-year writing courses
Jie Gao (Purdue University, USA), Zhaozhe Wang (Purdue University, USA),
Ge Lan (Purdue University, USA), Lindsey Macdonald (Purdue University,
USA) and Shelley Staples (University of Arizona, USA)

Citation practices and styles are integral to academic writing contexts. Previous research on citation use in academic writing shows important variability across citation form, function, and the stance of writers in relation to sources (Charles, 2006; Hyland, 2004; Swales, 2014). However, most studies have focused on advanced student and professional writing. Novice second language writers’ citation practices have not received the same level of attention in corpus-based studies (Keck, 2014; Pecorari, 2006), and they may require the use of different coding schemes. Using a corpus of L2 writing (majority L1 Chinese), we examine whether (1) the L2 writers’ citations vary in number across different assignments; (2) the writers show a preference for particular citation styles.

In this study, we consider two assignments from a first year writing course for L2 writers: (1) a literature review (LR) in which students closely examine sources and in (2) a research paper (RP) in which they use the same sources. Students registered for the writing course have TOEFL writing scores ranging from 19 to 27. Our 132 papers (147,000 words) for this study come from a larger L2 corpus comprising 7 million words from multiple sections of the same writing course. We calculated the number of citations and references in each assignment (per 1,000 words), and coded citations for integral/non-integral forms (Swales, 1990; Hyland, 2004).

Our preliminary results show that student writers used the same number of references in both assignments (32/1,000 words, about 3/paper) but more in-text citations in the literature review (LR=67; RP=45). Writers also used more integral than non-integral citations (73% integral: 79% LR; 64% RP). However, citation counts were highly variable, with 44% of writers using more than 10 citations and 33% using less than 3 per paper. The use of integral vs. non-integral citations also varied widely, with some writers showing a greater preference for non-integral citations. Compared to Swales (2014), who investigated upper level undergraduate and graduate student papers, we observed lower citation counts and references and opposite trends for integral/non-integral citation ratios (Swales found 73% non-integral vs. 27% integral in the MICUSP corpus).

These findings show that, on the whole, these L2 writers are still developing their citation practices, and such source work is extremely variable across assignments and individuals. The results have also exposed the need to develop a more detailed scheme to categorize the citation practices, as the classification of integral and non-integral citations does not seem to cover all the complexity and diversity of developing L2 writers’ choices. Our future work will draw on Charles (2006), who has labeled a third category as “general reference” in addition to integral and non-integral citation. O’Donnell (2008)
presented an annotation scheme which also provided new ideas for elaborating “non-integral” and “integral” criteria.

Our findings are expected to provide a more comprehensive analytical framework for understanding citation practices of developing L2 writers in the future. In addition, such work will undoubtedly open up greater opportunities for automatic citation identification and application to the teaching of developing L2 writers.

References


Testing usage-based theories with a representative corpus of nineteenth-century French
Angus B. Grieve-Smith (Columbia University, USA)

The researchers who compiled the corpus for the Trésor de la langue française dictionary in the 1960s, which became the FRANTEXT (2017) corpus that we use today, aimed to serve an audience drawn from “the upper and middle ranks of society” who desired to produce “careful enunciations, obeying empirical norms as much as they depart from belabored clauses” (Imbs 1971: XVIII). With that aim, they built the corpus around a “principle of authority,” examining literary histories of the nineteenth and twentieth centuries and collecting texts that were mentioned multiple times in those histories (ibid: XXIII).

FRANTEXT is large and well-made, which has earned it a place as one of the most popular corpora of the language. It appears to work well for helping its target audience produce their careful enunciations. But other corpus users, such as researchers in usage-based grammar, have needs that diverge widely from those of the audience of the Trésor de la langue française, and they would be better served with a different corpus design.

Usage-based theories, such as the theory of grammaticализация (e.g, Bybee and Thompson 1997), posit that the structures of a language have emerged from earlier states of that language. For example, when two constructions are in competition, the one with the higher type frequency - the greater mindshare - tends to increase in type frequency in subsequent years, while the other decreases. But the state of the language is not restricted to well-crafted works of literature as identified by later literary historians. Language users come from all social classes and have been influenced by spontaneous language, vernacular literature, and a kind of writing that is most lacking in literary corpora: bad literature.

To properly test usage-based theories we would need a corpus of every utterance that, for example, Alexandre Dumas fils was exposed to up to the time he wrote each of his plays. Collecting this data is beyond the abilities of science, but we can build a corpus that suits these purposes better than FRANTEXT. The Digital Parisian Stage Corpus aims to be that corpus.

The Digital Parisian Stage builds on an exhaustive catalog compiled by Charles Beaumont Wicks (1950 et seq.) of all plays that premiered in public in Paris in the nineteenth century, totaling over 30,000. The first phase of the project consists of a sample of thirty plays from the period between 1800 and 1815. This is a random one percent sample of the 2980 plays listed in Wicks’ first volume (1950).

Of this sample, the scripts for twenty-four plays have been obtained from Google Books, Gallica and other sources. Eighteen of them have been processed with Optical Character Recognition (OCR), and fifteen cleaned. Of those fifteen, three have been determined to be too short for most studies, leaving ten currently available for annotation.
The difference between the two corpora can be seen by brief investigations of well-known variables. For example, on average in the four theatrical texts for this period in FRANTEXT, 49% of negated declarative sentences used *ne...pas*, 21% *ne...point*, and 30% *ne* alone. In the twelve plays currently available from the Digital Parisian Stage Corpus, we find on average 75% *ne...pas*, 10% *ne...point* and 15% *ne* alone (*p* < 0.01).

A closer look at the texts can offer an explanation for this difference. Here is a quote from the play *Pinto* (Lemercier 1800):

(1) LA DUCHESSE: Ainsi votre esprit s'environne de tous les obstacles qu'il se crée; et si vous *n'en* aviez de véritables à surmonter, où serait la gloire de l'entreprise!

In the bolded section, the *ne* alone construction is used in a subordinate clause activating a presupposition, a pattern found in many earlier texts and one of the most common contexts for this construction. Contrast this with a quote from *le Grenadier de Louis XV* (Dubois 1815):

(2) ANSELME: C'est être bien hardi, après toutes les menaces que vous avez osé me faire si je *ne* vous donnais *pas* ma fille...

In this example, the *ne...pas* construction is used in a very similar context. It is the use of *ne...pas* in these contexts where *ne* alone had predominated that accounts for the much higher percentage of *ne...pas* in *le Grenadier de Louis XV* than in *Pinto*.

The higher proportion of *ne...pas* in the randomly selected plays likely due to the fact that three of the four FRANTEXT plays are dramas featuring aristocratic characters (like the Duchess in *Pinto*), while many of the randomly selected plays are vaudevilles and melodramas featuring characters who are servants, farmers (like Anselme in *le Grenadier de Louis XV*) and artisans. This in turn suggests that the playwrights believed these lower-class characters would sound more realistic with a higher proportion of *ne...pas*.

Since we know that there were far larger numbers of farmer, servants and artisans in early nineteenth-century France than there were duchesses and princes, we can surmise that the speech in *le Grenadier de Louis XV* is closer to what we might have heard in the countryside near Paris in that time - or in the time of Louis XV. But that play is not in FRANTEXT because there was nothing in *le Grenadier de Louis XV* to draw the attention of literary historians. If our goal is testing linguistic theories, we often have to go beyond the principle of authority.

Building on this success, work is progressing on three fronts. First, the list of thirty plays and the full text of the first fifteen plays were made available to the public on GitHub. Second, a new sample of thirty-one plays was drawn at random from Volume 2 of Wicks’ catalog (1953) and the process of obtaining, OCR and cleaning has begun for that sample. Third, additional linguistic variables are being selected and annotated in the corpus.
References


How Data-Driven Learning Can Help EFL Learners Improve Their Micro Level Skills of Writing
Mehrdad Sepehri (Islamic Azad University, Shahrekord Branch, Iran)

This study was conducted to investigate the effects of Data Driven Learning (DDL) on EFL learners' writing skills development as far as their 'Micro level skills' is concerned. 'Writing skills' in the present study are confined to the formal or structural aspects of the language, the way the linguistic elements, words, phrases, clauses, and sentences form larger units of language to transfer the ideas and concepts. The objectives of the study were to compare the learning effects of DDL method with the conventional teaching method's effects on the measures of 1) learners' declarative knowledge of the taught materials, and 2) analytic scoring of their written products.

A pretest posttest control group design was employed to collect the required data. Two groups of EFL university students attending a 'Paragraph Development' course were compared as far as their micro-level writing skills are concerned. The control group consisting of 24 male and female EFL learners received instructions through conventional method of textbook usage, teacher explanations and classroom exercises.

The experimental (DDL) group including 26 male and female EFL learners received a certain number of classroom concordance-based handouts in addition to textbook usage accompanying with shorter teacher explanations and partial class activities. In order to prepare the handouts the two learner corpora of ICLE and IrCLE (particularly compiled for this study) were used as the main sources of concordancing lines. Using the two corpora and the AntConc 3.2.4w software (Anthony 2012) and based on the 'Paragraph Development' course syllabus, and the textbook analysis conducted to determine the target linguistic elements of the course, a series of DDL-based units were prepared. These handouts were used at the beginning of each class session for 20 to 30 minutes. The rest of the time in each session was allocated to work with the same textbook as used in the control (Non-DDL) group. In addition to textbook usage, both groups were asked to do similar weekly assignments, and attended the same pretest and posttest examination sessions, on the second and final term sessions respectively. Finally, their performances on the pre- and posttest were analyzed quantitatively in order to answer the research questions:

1) Do Iranian EFL learners achieve higher improvements in their declarative writing scores after being taught through the DDL method than those who were taught through a Non-DDL method?
2) Do Iranian EFL learners achieve higher improvements in their analytic writing scores after being taught through the DDL method than those who were taught through a Non-DDL method?
Statistical analyses conducted on different parts of the pre- and post-test showed that there was a significant difference between the two groups in terms of declarative knowledge, the 'knowledge stored as facts' (Ellis, 1994, p32) or as defined by Anderson (1983) the 'knowledge that consists of factual information about the L2 that has not yet been integrated or automatized'. This declarative knowledge was tested through the first 5 parts of the pre- and post-tests in which the learners were asked to show their familiarity with linguistic structures, connectors, and phrases frequently used in developing a paragraph. The results showed that class attendance with the DDL-based materials has had more positive effects on the learners' improvement than a conventional one with typical textbook.

As for the procedural knowledge of the learners' language development, an 'Analytic Scoring' framework was used to compare the written products of the two groups. The Analytic Scoring scale, developed by Jacobs et al. (1981), was revised in a way in which different aspects of writing (Content, Organization, Vocabulary, and Language Use) were rated. 'Content' deals with the extent to which the writer is knowledgeable about the subject matter and the topic of the task. 'Organization' refers to how clearly the ideas are sequenced, supported and developed in a piece of writing. 'Vocabulary' indicates the proficiency level of the writer in word knowledge, range of idiom choice and usage as well as appropriate word form. And finally 'Language use' relates to the ability of the writer in making use of grammatical structures, agreement, tense, number and word order/function, articles, pronouns and prepositions at the sentence level. In order to find a reasonable answer to the second question appropriate statistical analyses were conducted.

The results showed that both groups have improved an overall extent of the features of this measuring scale. Since no significant difference between groups' improvements was observed in regard to the Analytic scoring as a whole (at the macro-level comparison), it can be concluded that the observed improvements in the total Analytic scores in both groups might not be attributed to any particular method of teaching, although instructional activities in both groups have been effective. Detailed analyses revealed that there was no statistically significant difference between the two groups in terms of 'Content', 'Vocabulary', and 'Organisation'. However, the DDL group demonstrated more improvement in 'Language use' indicating that the DDL-based units have given the learners an advantage in learning and applying the target grammar patterns. In other words, the DDL-based units can be seen to act as supportive instructional materials in developing the components of the 'language use' feature.

The implications of the study have been examined on the two areas of theoretical and practical. As for theoretical implications, some underlying ideas of DDL such as authenticity of materials, discovery learning, and the noticing hypothesis have gone under further scrutiny. Using corpora in language instruction, and applying concordancing in preparing language teaching materials are not the only but among the practical implications of the study.
References


A corpus of contemporary Czech poetry

Michal Škrabal (Institute of the Czech National Corpus, Charles University, Czech Republic), Karel Piorecký (Institute of Czech Literature of the Czech Academy of Sciences, Czech Republic), Petr Plecháč (Institute of Czech Literature of the Czech Academy of Sciences, Czech Republic), Robert Kolár (Institute of Czech Literature of the Czech Academy of Sciences, Czech Republic), Petr Kuběnský (Institute of Czech Literature and Library Studies, Masaryk University, Czech Republic) and Pavel Procházka (Institute of the Czech National Corpus, Charles University, Czech Republic)

The aim of this joint project of two academic institutions in the Czech Republic – Institute of Czech Literature and Institute of the Czech National Corpus – is to build a corpus of contemporary poetry (cf. Plecháč & Kolár, 2015) that would, with the methods of corpus linguistics and also within the methodological intentions of digital literary studies (cf. e.g. Jockers, 2013; Moretti, 2013; Piorecký, 2015, 2016), investigate Czech language poetry texts published in their first editions since 1990 and going up to the present time (we have set the year 2015 as a boundary line). The goal is, above all, to create a broad empirical foundation on whose basis it shall be possible to identify and verify the specifics of contemporary poetic language and to strengthen the literary-historical knowledge of Czech poetry of the recent past and of the present, thus in the period when it was significantly influenced by a change in the media paradigm given by the arrival of digital media.

In the corpus there are to be found texts compromising, thanks to their elite artistic level, the notional centre of the literary system (its poetic subsystem): especially books published by respected publishers, which follow the standard editorial process and whose edition profile is built on the rigorously chosen principle. A second segment of the corpus are texts situated outside of the demarcated centre of the literary system – primarily because of the not matter-of-course literary competence of their creators. This will above all concern a wide field of texts published on open internet platforms, at which editorial selection and the processing of texts before they are presented to the public is eliminated (so called literary servers, blogs, etc., but also some self-published titles). The representativeness of the corpus should not be secured by the canonization method (by the choice of what is ostensibly the best) but, to the contrary, by a wide base of sources collecting a relatively exhaustive typological and qualitative spectrum of contemporary poetry (from representative publishers to amateur web pages).

Besides the creation of a database for the investigation of literary language, the project would also focus on research questions which it is possible with the collection of the data to solve. Also of significance is the interdisciplinary nature of the entire project: the quantity of themes is mutual both for linguists (e.g., the common lexical characteristics of poetic language, especially in contrast to non-poetic language, and the specifics of lexical semantics; also potentially the problem of the metaphor, poetisms, etc.) and for literary studies (e.g., versification...
characteristics – cf. Plecháč, 2016, thematic assemblages and genre structures of poetry production in individual types of media).

References


Extracting construction networks from Cantonese speech corpora using clustering algorithms
Andreas Liesenfeld (Nanyang Technological University, Singapore)

Language is constructions “all the way down”, concludes Adele Goldberg (2006:18). Starting from this key insight associated with construction grammar (CxG), this study addresses an issue that is a result of CxG’s assumption that “the network of constructions captures our grammatical knowledge in toto” (ibid): the nature of the network organization of constructions.

Adopting a usage-based constructionist approach, this ongoing PhD-level study aims to model structural properties of construction networks through clustering properties of in-situ speech sequences. This data-driven grammar induction draws on recent developments in CxG that conceptualize language as complex adaptive systems (CAS) (Beckner et al. 2009). The Language-as-CAS approach holds that construction networks emerge from interrelated patterns of social interaction, experience and cognitive processes. Grounded in natural speech data, the study aims to provide empirical evidence for this emergence process, exploring ways of how construction networks can be identified, extracted and presented as complex adaptive system networks.

The goal of the project is to explore ways of how typologies of constructions can be extracted from Cantonese speech data organized in the CHAT format (minCHAT) (MacWhinney 2000). Utilizing various clustering algorithms, the subsequent network extraction is based on three key mechanisms that shape construction networks identified by Ellis (2012): frequency, recency and context. Preliminary results of this data analysis show how structural properties of speech corpus data can be extracted by applying clustering algorithms that simulate domain-general cognitive constraints in humans. The extracted construction networks can be used to complement or replace existing Cantonese grammar formalisms for various NLP tasks, such as natural language understanding, and provide a new model for data-driven grammar induction.

References


Ellis, N. C. (2012). What can we count in language, and what counts in language acquisition, cognition, and use. Frequency effects in language learning and processing, 1, 7--34.


Recent years have seen a surge of interest in learner corpus research (e.g. Granger, Gilquin, & Meunier, 2015; Paquot & Granger, 2012). Much of this research has focused on the use and development of vocabulary, both single words and collocations. However, much of the work has been done with learner data collected at one point in time. Longitudinal studies are still rare, such that researchers have called for a greater emphasis on studies conducted over a period of time with the same group of learners (e.g. Laufer & Waldman, 2011; Paquot & Granger, 2012). In addition, many of the current studies are limited in a number of ways. First, researchers have for most part focused on upper intermediate, or advanced learners. How the many and varied aspects of vocabulary use (e.g. the use of single words, collocational knowledge development, lexical diversity, etc.) develop in less proficient learners is still poorly understood. Second, the topics in the groups being compared (e.g. less vs. more proficient second language [L2] learners, native vs. non-native writers, etc.) have often not been controlled for. This could have potentially introduced a confound affecting the results reported. Third, most of the longitudinal studies that have looked at production of L2 collocation have employed only a handful of participants (e.g. Crossley & Salsbury, 2011; Li, Eskildsen, & Cadierno, 2014; Li & Schmitt, 2009; Yuldashev, Fernandez, & Thorne, 2013; but see Siyanova-Chanturia, 2015). Thus, most of the studies looking at collocation use and development have been case studies or cross-sectional studies, potentially limiting our understanding of the learning process. In addition, it is noteworthy, that the majority of learner corpus studies with a focus on vocabulary have looked at English as a L2. Relatively few studies have investigated L2s other than English.

To address the above concerns in the current learner corpus research focusing on vocabulary, a large-scale\(^1\) longitudinal corpus of L2 Italian (first language [L1] Chinese) was collected. In total, 175 learners contributed two essays to the corpus. One essay was written at the beginning of a six-month course of Italian, and the other was written at the end of the course. The students were enrolled in a full-time course of Italian as a second language, which took place at a university in central Italy. Students of three proficiency levels contributed to the corpus: A1 (n=39), A2 (n=86), and B1 (n=50). All students came from China and were between 17 and 33 years of age.

\(^1\) Large-scale in terms of the number of learners who participated in this longitudinal study. While the resulting corpus is relatively small, the large number of learners (n=175) by far surpasses participant pools used in earlier longitudinal learner-corpus studies.
(mean=20.5, SD=2.7; 105 females). On average, the students spent 1.7 months in Italy (range 0.5-5, SD=0.69) prior to writing the first essay. The exact same 175 students who wrote the first essay also wrote the second essay. The students who only wrote one of the two essays were not included in the corpus. Three similar essay topics were offered: 1) My first impression of Italy and Italians, 2) My hobbies: what do I usually do in my free time, 3) My last holidays. The students were instructed not to write on the same topic more than once. Hence, all students chose two of the three topics. Finally, the students were taught by the same group of teachers at the same university. Thus, it can be said that the corpus creation addressed the issues of topics, teaching style and learning environment. Importantly, a range of proficiency levels is represented in the corpus: A1, A2, and B1 (according to CEFR). The total size of the corpus is circa 97,000 words (data collection Level 1: A1=7,126, A2=22,851, B1=15,903; data collection Level 2: A1=9,487, A2=24,117, B1=17,386).

The main aim of the present investigation was to examine vocabulary use – in terms of single words and collocations – in essays collected at the beginning of the course (Level 1) versus those collected at the end of the course (Level 2), separately for the three proficiency levels. Thus, we were interested in the progress (if any) for A1 learners at Level 1 vs. Level 2, A2 learners at Level 1 vs. Level 2, and B1 learners at Level 1 vs. Level 2. To this aim, a number of analyses were conducted.

First, the POS distribution (nouns, adjectives, adverbs, and verbs) was analysed. The use of the POS was found to be comparable with the exception of nouns. Level 1 essays were found to contain significantly more nouns than Level 2 essays for A1 and A2 writers (but not B1).

Second, we looked at the number of tokens per essay and the number of tokens per sentence in Level 1 essays versus Level 2 essays. Level 2 essays were found to contain significantly more tokens than Level 1 essays in A1 and B1 writers (but not A2).

Third, we used the Guiraud index (e.g. Guiraud, 1954) as an index of lexical diversity. This index was used instead of type/token ratio because it compensates the systematical decrease of the number of tokens when texts to compare have different lengths (e.g. Van Hout & Vermeer, 2007). Level 2 essays were found to have a consistently higher Guiraud index. The differences between Level 1 vs. Level 2 were found significant across all three proficiency levels, but were particularly prominent in A1 and B1 writers.

The above results suggest that Level 1 and Level 2 essays appear to differ in terms of lexical diversity and the students’ ability to produce longer pieces of writing, rather than in terms of POS distribution. This trend is particularly and consistently noticeable in beginner (A1) and intermediate learners (B1).

Finally, we looked at L2 learner use of collocations of different types (e.g. noun+adjective). To this aim, all learner items were first extracted

---

2 It needs to be noted that no learner advanced from one CEFR level to the next between data collection points. So, all A1 learners remained A1, A2 remained A2, and B1 remained B1.
automatically from the learner corpus. A L1 reference corpus – Paisà (Lyding et al., 2014) – was used to extract frequencies of learner items in a representative corpus of L1 Italian. These frequencies were also used to calculate measures of association strength: t-score and mutual information. For the data analysis, we opted for mixed-effects modelling (Baayen, Davidson, & Bates, 2008). Preliminary analysis suggested a relatively comparable collocation usage, in terms of L1 frequencies and measures of association strength, for Level 1 vs. Level 2 across the three proficiency levels. It appears that while six months were sufficient for the students to be able to produce longer pieces of writing and to exhibit greater lexical diversity, this period of time might have not been sufficient for these learners to improve substantially in their use of L2 collocation (although some improved was observed).

In sum, the present large-scale longitudinal investigation of L2 Italian writing has provided a rich picture of how vocabulary use evolves in an Italian as a second language environment. Despite the relative brevity of the course – only six months – we observed clear developmental patterns, as suggested, for example, by longer sentences and essays, as well as higher lexical diversity indices. In line with previous research, however, our analysis also showed that L2 collocation learning can be slow and uneven (e.g. Laufer & Waldman, 2011).

References


The English Grammar Profile (EGP) Project was a four-year quasi-longitudinal study investigating learner grammar from the Cambridge Learner Corpus (a 55 million-word learner corpus containing over 200,000 exam scripts, from over 200 countries, by speakers of over 140 first languages). The focus of the project was to scrutinize the grammar used by learners in the Cambridge English Language Assessment exams across the levels of the Common European Framework of Reference (CEFR), tracking the evolution of the use of forms and features, across different learners, contexts, L2 backgrounds, levels of proficiency. The main output of the research is the EGP, a free educational online database, which provides a profile of over 1,200 corpus-based grammar competency statements about real learner grammar use across the six CEFR levels.

The main focus of this paper will be on the findings of this extensive corpus study in relation to the acquisition of grammar across different levels of competency, as well as interesting insights into our misconceptions about learner grammar and its development across competency levels. These findings parallel and/or complement existing corpus and non-corpus SLA work.

We note that there are some important departures in this study from hitherto pioneering contrastive learner corpus research on interlanguage:

By using the CEFR as the calibration, we are looking at learner language across all six staging proficiency points A1-C2, which have been intuitively drawn within the CEFR (Council of Europe 2001).

1. By using the CEFR as the calibration, we are looking at learner language across all six staging proficiency points A1-C2, which have been drawn within the CEFR (Council of Europe 2001). Hitherto studies have looked at only some levels.
2. We are investigating a summative aggregation of grammatical competences of learner English (from those who have taken Cambridge English exams). Hitherto studies usually examine one or a set of grammatical features.
3. We do not take a systematic contrastive focus between learners’ L1 and the target language.
4. We are not setting out to examine the learner English of one L1 cohort; we are exploring data from a multitude of L1 backgrounds across 55 million words of learner data.
5. We are not tracking the evolution of any one interlanguage feature.
6. We are investigating the quasi-longitudinal development of grammar competence, with data spanning 17 years, across more than 260,000 learner exam documents.
We are not setting out to chart learner development in relation to error decline, plateau or regression. We are tracking and describing the development of learner grammar competence towards an 'idealized L1 norm'.

The authors acknowledge the potential circularity in the use of exam data from learners who will have been preparing for the exams using targeted exam practice material. We anticipate that learners will have been exposed to a range of input, pedagogical or otherwise, other than traditional ELT textbooks and exam-focused materials.

As noted by Ortega & Byrnes (2008), accuracy development at the higher levels of proficiency has been under-researched and this is echoed by Thewissen (2013) who makes her own substantial empirical contribution by looking at learner error data developmentally, within the CEFR framework. Interestingly, there is complementarity between Thewissen's (2013) findings on grammar errors across levels of the CEFR and the patterns of EGP grammatical competencies. Thewissen (2013), in referring to instances where there is lack of significant progress in accuracy between B2 and C2 levels (e.g. lexical phrases), cites the accuracy-complexity trade off effect whereby learners at higher levels take more risks with language and use it in more complex ways but in doing so their risk of making an error increases and so this can hinder improvement in accuracy. This stabilization, depending on where it occurs, according to Thewissen, is related to the “ceiling effect” (after Milton & Meara 1995), which could be at play between B2 and C2. She advises that although errors still remain, a significant amount of learning has already taken place and she stresses that the stabilization tendency should not be narrowly interpreted as an absence of any development whatsoever, or in Larsen-Freeman's words “linguistic rigor mortis” (2006: 597).

The present study of competence, not error, brings some insight to the ceiling effect. We often see a ‘syntactic ceiling’ where a form has stabilised and reached its syntactic ‘developmental endpoint’ (after Thewissen 2013) at A and B levels but as learners go beyond these levels, they deploy the stabilised forms with a greater complexity of meaning and with greater dexterity of use. As they get to higher levels, this is constantly borne out in how a learner can use a syntactic form pragmatically and skillfully play with a form for greater subtlety of meaning or focus.

The past simple tense offers one of many examples of this developing functional and pragmatic competence. At A1, the affirmative form is evident and stable, but only with a very limited range of regular and irregular verbs, leading to the can-do statement ‘Can use the affirmative form with a limited range of regular and irregular verbs'. The only function of the form at A1 that meets the criteria for competent use (based on the methodology for this research) is to talk about everyday events or states, such as Every time I went to Hendon Park, I saw so many people there. (A1, 2007, Polish).

This study illustrates that two developments happen as learners go up levels of the CEFR:

1) they learn more verbs and so their lexical range of verbs expands, and
2) the past simple affirmative form is deployed in more complex ways, illustrating growth in functions and often pragmatic competence.
Were we to look at the pattern of errors, as others have done, we would not see this picture of growing competence beyond point of stabilisation of form. We would rather see a ceiling effect, a stabilisation or developmental end point at A1 in the correct use of the affirmative form and miss a rich seam of competence from A1 upwards. Put simply, our insight is that by looking at competence rather than error stabilisation end points, we can see how learners develop beyond the acquisition of a form to how they can use that form with greater functional and pragmatic effect and complexity.

Thewissen (2013) rightly points to the dearth of attention to the importance of investigating lexical development, apart from work such as Ife et al (2000) and Milton & Meara (1995). Our findings offer complementary insight into the role of lexis. As illustrated by past simple affirmative form examples at A1 level above, while the syntactic form can be used competently at A1, it can only be used with a limited range of verbs. This is because a learner at A1 has not yet acquired many verbs. In the present study, we see an increase in lexical repertoire going hand-in-hand with an expanding repertoire of grammatical uses of the same form, level by level. Certain verbs are more typically used for particular functions and so as a learner’s lexical range increases so does the potential range of usage contexts increase.

Within the ELT community there is widespread agreement on what grammar structures are taught and when. Other insights from this study include some mismatches between what is taught at certain levels in a traditional ELT syllabi and when it is actually competently acquired. In some cases, we found more frequent use of structures in the learner data than in native speaker corpora. In other cases, we even found that learners showed competency in the use of grammatical items before they appear formally on traditional syllabi. This may be due to a number of factors, some of which are related to individual differences and variation, or indeed, it may link to the findings from usage-based theories of learning (Ellis et al. 2015, Gablasova et al. 2017).

Learner use of conditional clauses illustrates an example of competence of structures not formally on the syllabus. Students are traditionally taught four conditional clauses, described as zero, first, second and third conditionals, whose forms are limited to four prescribed patterns. The four conditional clauses are typically introduced one by one from A2 level to B2. Two surprises emerge from the learner data:

1) evidence of all four traditional clauses in stable use by B1, and
2) evidence of 27 different forms and uses expressing conditionality from A1 to C2 (supporting the findings of Gabrielatos 2010).

It appears that learners are showing implicit knowledge of forms that have been both explicitly and not explicitly taught in the classroom. Here we find resonance with interest in explicit and implicit learning and theories of incidental vs intentional language learning.

In summary, by looking at a 55 million word corpus of learner English exam writing, from a global range of L1 backgrounds, we have been able to explore largescale grammar patterns of use and their development across the CEFR. The
insights from this research add to the existing body of work on learner corpora but do so from the perspective of what learners can do rather than what they cannot.

References


Report on the Automatic Extraction of Korean Scientific Phrasal Term Candidates, with a Focus on Science Textbook Corpus

Jun Choi, Hae-Yun Jung, Hyeonah Kang and Seiyeon Kim
(Kyungpook National University, Korea)

In 2011, the International Association for the Evaluation of Educational Achievement (IEA) conducted the fifth ‘Trends in International Mathematics and Science Study’ (TIMSS 2011), in which fourth-grade students from 50 countries and eighth-grade students from 42 countries participated. The results of TIMSS 2011 revealed an interesting fact about Korean students. While they did achieve high ranks in scientific subject assessments, the rates of their interest in this field showed nonetheless the lowest scores.

Fang (2006) and Wellington & Osborne (2001) have suggested that the difficulty of scientific terminology was one of the major reasons why students find sciences hard and tend to avoid scientific studies. It seems in fact that scientific terminology constitutes a determining factor in students’ interest in and understanding of scientific subjects. In that sense, we have been carrying out a project titled ‘Development project of an integrated search system of educational science terminology through the construction of science terminology database’. In this paper, we present the procedure for automatically extracting scientific phrasal term candidates, which is part of the science textbook corpus building and term annotation processes, and report on the results of the science specialists and science education experts’ examination of the list of candidates.

The list of morphosyntactic patterns (POS-gram) of scientific phrasal terms is compiled in two phases. A first list is obtained by analyzing the patterns of the 220,000 phrasal terms that are included in existing language and terminology dictionaries. Secondly, the list is refined by examining and identifying the conventional usage in the science textbook corpus during the morphological annotation process. This allows us to extract 437 types of POS-gram. This science textbook corpus consists of 6,000,000 words and is based on 132 science textbooks for grade 1 to grade 12 (i.e., for primary, middle, and high schools), used from 1992 to date.

The final list goes through a process of elimination so as to compile a list of stopwords, on the basis of which the POS-gram computation generates a list of automatically extracted candidates, which is then handed on to science specialists and science education experts for terminological annotation. Whenever a form of that list of candidates matches the one given by the dictionary, the respective specialised field information is assigned to it; the other terms are differentiated without any particular information being assigned. Finally, we discuss the issues related to the various scientific experts’ examination and assess to what extent the automatic extraction of scientific phrasal term candidates is accurate and effective. As we investigated the feasibility of the automatic annotation of terms based on the morphological patterns in a given field, we could redefine the characteristics of that field and verify the emergence of new terms, thereby contributing to improving communication within the field.
References

How large is the BNC? A proposal for standardised tokenization and word counting
Vaclav Brezina and Matt Timperley (Lancaster University, UK)

Introduction

One of the core principles of the scientific method is replicability of results. Observations and experiments need to be exactly repeatable with different datasets in order to establish the stability of the findings (e.g. Asendorpf, 2013). Corpus linguistics aspires to being a scientific approach to language analysis and thus needs to satisfy the replicability requirement. This primarily involves standardisation of the procedures and instruments used in the field (McEnery & Hardy 2011).

Replicability in corpus linguistics is a fundamental issue which, however, has received only limited attention. For example, the differences in language use between multiple corpora sampling the same type of language have been explored in Gablasova et al. (2017) showing a large amount of variation between these corpora. We aim to take this research one step further to investigate the effect of different instruments with the same dataset. As an example dataset, we used the British National Corpus (BNC), which is a widely used dataset in corpus linguistic research, and six corpus tools commonly used for the analysis of the BNC.

In this experiment, we observe both i) variation in the overall token counts given by different tokenization procedures implemented by the tools as well as ii) variation in the frequency counts of individual linguistic variables, which is also connected to different ways of identifying tokens in the corpus. The main purpose of the experiment is to empirically analyse the amount of variation observed and assess its impact on the replicability of results in corpus linguistic studies. It is important to realise that the variation that we observe in this study is purely methodological and results from employing different tools. This type of variation is thus highly problematic because it shows that results produced by different tools are not comparable. The results therefore cannot be used easily in e.g. meta-studies, which bring together the knowledge in the field.

As mentioned earlier, the source of the variation in our experiment is the tokenization procedure and the lack of standardisation in this area. Tokenization is the process of splitting text into atomic parts. In segmented languages, like English, this is often considered an easy task because it can rely on graphical clues in the texts (spaces, punctuation etc.). This gives rise to a so called ‘graphic word’ as a unit (atomic part). Kučera and Francis (1967: 3) define a graphic word as “a string of contiguous alphanumeric characters with space on either side; may include hyphens and apostrophes but no other punctuation marks”. The graphic word, however, is not the only definition employed in tokenizers for corpus tools; additional tokenization criteria are applied to deal with particular languages and ambiguous cases. These are, however, rarely openly stated.

The difficulties with tokenization generally arise in the following cases:
1. **Non-segmented languages.** Languages such as Chinese do not have delimiters to denote separation between words. This needs to be added using a language-specific process called segmentation.

2. **Punctuation.** Different dashes/hyphens can indicate that two strings of characters should be considered one, or, in other cases, two units, e.g. a forget-me-nots vs. a word-a string of characters-which... An apostrophe can indicate the end of an open quote, possessive case (e.g. Peter’s) or a contraction (C’mas).

3. **Clitics.** Different decisions can be made about clitics and their status. Clitics such as ’ll can be treated as atomic units or can be considered a part of larger atoms like we’ll.

4. **Abbreviations.** Full stops may not occur with graphic words but might be considered an atomic unit when part of an abbreviation e.g. in this example.

5. **Multi-word expressions.** This overlaps with clitics. For some purposes, idioms or other multi-word expressions may be considered atomic (Webster and Kit, 1992).

6. **Data noise.** If the corpus texts were obtained via e.g. Optical Character Recognition (OCR) then characters may be mistaken for others. This can lead to situations that violate a static definition of an atomic unit.

The difficulties of tokenization are well presented in Yamashita and Matsumoto (2000), who propose a method of treating segmented and non-segmented languages with a single tokenizer. However, currently, there is no single tokenizer, or set of identical principles, used for all corpus tools. This means that when counting tokens there will be discrepancies between tools. The ambiguities might skew the counts in only a small proportion of cases, but this can be inflated by corpus size. We continue by investigating the impact of decisions taken at the tokenization stage on the analysis of the BNC.

**Method**

The BNC XML version, one identical dataset, was used in different corpus tools. The impact of different tokenization procedures on the searches was explored. Six different software tools were used: CQPWeb, BNCweb, BNC-BYU, Sketch Engine, Xaira and #LancBox.

**Results**

The results show that the overall token counts for the same dataset, the BNC, vary from 96,263,399 in the BYU interface to 112,289,776 in the Sketch Engine. The main source of the variation is the decision to include or exclude punctuation as token counts.

<table>
<thead>
<tr>
<th></th>
<th>CPQWeb</th>
<th>BNCweb</th>
<th>BYU</th>
<th>SkE</th>
<th>Xaira</th>
<th>#LancBox</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tokens</td>
<td>112,102,325</td>
<td>98,313,429</td>
<td>96,263,399</td>
<td>112,289,776/112,181,015 (CLAWS)</td>
<td>112,532,992</td>
<td>96,960,485</td>
</tr>
<tr>
<td>Words (if different)</td>
<td></td>
<td></td>
<td>96,133,793/96,052,598</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Tokenization does not only affect the total word counts in the corpus and normalization of the data, it also plays a role in searches for individual linguistic variables as is apparent from Table 2 below.

<table>
<thead>
<tr>
<th>Tools</th>
<th>CPQWeb</th>
<th>BNCweb</th>
<th>BYU</th>
<th>SkE</th>
<th>Xaira</th>
<th>LancBox</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>the</em></td>
<td>6,041,234</td>
<td>6,041,234</td>
<td>5,971,799</td>
<td>6,054,939</td>
<td>6,055,159</td>
<td>6,054,559</td>
</tr>
<tr>
<td><em>the</em> per 1M</td>
<td>53,890.35</td>
<td>61,448.72</td>
<td>62,036.03</td>
<td>53,922.40</td>
<td>53,807.86</td>
<td>62,443.57</td>
</tr>
<tr>
<td><em>new</em></td>
<td>124,022</td>
<td>124,022</td>
<td>121,881</td>
<td>124,399</td>
<td>124,308</td>
<td>124,235</td>
</tr>
<tr>
<td><em>new</em> per 1M</td>
<td>1,106.33</td>
<td>1,261.5</td>
<td>1,266.12</td>
<td>1,107.80</td>
<td>1,104.64</td>
<td>1,281.30</td>
</tr>
<tr>
<td><em>research</em></td>
<td>26,682</td>
<td>26,682</td>
<td>26,566</td>
<td>26,702</td>
<td>26,793</td>
<td>26,692</td>
</tr>
<tr>
<td><em>research</em> per 1M</td>
<td>238.01</td>
<td>271.4</td>
<td>275.97</td>
<td>237.80</td>
<td>238.09</td>
<td>275.29</td>
</tr>
<tr>
<td><em>mauve</em></td>
<td>214</td>
<td>214</td>
<td>213</td>
<td>208</td>
<td>214</td>
<td>206</td>
</tr>
<tr>
<td><em>mauve</em> per 1M</td>
<td>1.91</td>
<td>2.18</td>
<td>2.21</td>
<td>1.85</td>
<td>1.90</td>
<td>2.12</td>
</tr>
</tbody>
</table>

**Table 2.** Frequencies of linguistic variables in the BNC according to different tools

Focusing on the raw frequencies (first entry row for each linguistic variable), we can see that in the case of the, the most frequent lexical item in English, the frequency counts range from 5,971,799 in the BYU platform to 6,055,159 in Xaira; this is a 1.4% difference. In the case of new and research, two mid-frequency items, the differences between the smallest and the largest frequency count are 2.07% and 0.85% respectively. Finally, the low-frequency item mauve shows a difference of 3.9%. The relative (normalised) frequencies (counts per 1M tokens), which combine the effect of total token counts for the BNC in the individual tools and the frequency counts for the individual variables, indicate even larger discrepancies among the individual tools.

**Conclusion: Proposal for more rigorous analyses**

Having shown the implications of different tokenization principles implemented in different widely-used corpus tools and having demonstrated the effect of these on the results of quantitative analyses in corpus linguistics, we propose two innovations: i) Tokenization Parameters Notation (TPN) and ii) LOB as a unit of corpus size measurement. First, we introduce the Tokenization Parameters Notation (TPN), which uniquely describes the principles of tokenization implemented in a corpus tool. Second, we introduce LOB, a new unit of corpus measurement and normalization, which is independent of the individual corpus tokenization. LOB is defined as the size of the Lancaster-Oslo/Bergen corpus (LOB) given the particular tokenization principles. It is approximately one million tokens but can vary across different corpus tools. The advantage of using LOB-based normalization, as opposed
to normalization to a particular fixed basis, e.g. per million, is that it allows larger comparability of results based on different tools. 5

References


Creating a Bespoke Corpus Sampling Frame for a Minoritised Language: CorCenCC, the National Corpus of Contemporary Welsh

Mair Rees (Swansea University, UK), Gareth Watkins (Cardiff University, UK), Jennifer Needs (Swansea University, UK), Steve Morris (Swansea University, UK) and Dawn Knight (Cardiff University, UK)

Overview

In this paper, we discuss the steps taken to create a bespoke sampling frame to use when constructing a national corpus for a minoritised language, in this instance Welsh. We illustrate the processes involved in designing this sampling frame, with emphasis on how decisions were reached to reflect this context and how they are different to those which might be made in the construction of a sampling frame for a majority language such as English. The processes adhered to, and the decisions made as part of this design process, are potentially of value to other corpus linguists who may be interested in creating corpora for other minoritised languages.

As is widely documented, the design and construction principles used to build corpora are often locally determined (Conrad, 2002: 77), depending on the vision for the corpus and the questions that it is intended to help us answer. CorCenCC (Corpws Cenedlaethol Cymraeg Cyfoes – The National Corpus of Contemporary Welsh) is an interdisciplinary, collaborative project, whose vision is to construct the first general principled corpus of contemporary Welsh.

CorCenCC will contain 10 million words by the end of the project, comprising 4 million each from spoken and written sources and 2 million from digitally mediated sources (e-language). Data will be drawn from a variety of contexts, ranging from formal (e.g. political documents, televised interviews and formal letters) to less formal ones (e.g. diaries, phone calls and text messages). A provisional outline for the distribution of this data is shown in Figure 1.

This paper will examine some key issues, questions and considerations involved in creating a bespoke sampling frame for the Welsh language.

Spoken language

The strategy used to collect spoken data for CorCenCC is adapted from CANCODE’s ‘genre’ approach, which ‘tries to seek a balance between speaker, environment, context and recurrent features’ (McCarthy, 1998: 8), in combination with the BNC’s ‘context-governed’ model (Crowdy, 1993: 259). CANCODE includes data from five contexts – Transactional, Professional, Pedagogical, Socialising and Intimate (‘Private’ in CorCenCC’s terms) – but omits contexts associated with more formal language due to its focus on ‘unrehearsed, non-formal talk’ (McCarthy, 1998: 9). As a general corpus, CorCenCC does not have the same focus as CANCODE. It is important, therefore that both formal and informal language are represented within the corpus, particularly given that the syntax and morphology of Welsh vary considerably according to degrees of formality. For this reason, two contexts from the BNC’s model – Public/Institutional and Leisure (‘Media’ in CorCenCC’s terms) – have been added to
CorCenCC’s spoken sampling frame, allowing it to reflect more fully the complexity of the Welsh language.

<table>
<thead>
<tr>
<th>Cyd-destun / Context</th>
<th>%</th>
<th>Geiriau / Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cyhoeddus/Sefydliadol / Public / Institutional</td>
<td>10%</td>
<td>400,000</td>
</tr>
<tr>
<td>Cyfrydau / Media</td>
<td>15%</td>
<td>600,000</td>
</tr>
<tr>
<td>Trafodol / Transactional</td>
<td>10%</td>
<td>400,000</td>
</tr>
<tr>
<td>Proffesiynol / Professional</td>
<td>10%</td>
<td>400,000</td>
</tr>
<tr>
<td>Pedagogaid / Pedagogical</td>
<td>10%</td>
<td>400,000</td>
</tr>
<tr>
<td>Cymdeithasu / Socialising</td>
<td>22.5%</td>
<td>900,000</td>
</tr>
<tr>
<td>Preifat / Private</td>
<td>22.5%</td>
<td>900,000</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>100%</td>
<td><strong>4,000,000</strong></td>
</tr>
</tbody>
</table>

**Figure 1:** Provisional sampling frame for CorCenCC

Sampling from the seven contexts shown in Figure 1 will enable CorCenCC to represent all the different genres that are spoken and heard, including the language of the media, public and institutional language, the language of the workplace and of education, as well as personal and social use of Welsh. Data will also be collected from speakers from all regions of Wales, of all ages and genders, with a wide range of occupations, and with a variety of linguistic backgrounds (e.g. how they came to speak Welsh), to reflect Wales’ diversity not only of genres but also of Welsh speakers themselves.

**Written language**

CorCenCC’s written sub-corpus is constructed on a genre-based framework in order to produce an ‘information-rich, user-friendly resource’ as outlined in Lee (2001: 63). The key challenges (at sociolinguistic, socio-political and practical levels) in compiling a sampling frame for written language, a sampling frame which is both balanced and representative of the Welsh language community, are illustrated in this paper. Firstly, there is a relatively disproportionate number of children’s books which are published each year, for both demographic and commercial reasons (Rosser, 2012: 1). Secondly, the complexities of accurately representing the range and distribution of literature for
adult and child second-language learners at various levels of expertise. Thirdly, are the broader issues around readership, authorship, demography and dialect which again have impact on the balance and representativeness of the corpus.

These challenges were considered when designing the sampling frame for the written sub-corpus. For example, the medium ‘Books’ includes books written for adult learners, children, and child learners. It is anticipated that this will be of assistance to pedagogical end-users in the future, as these sub-sets will demonstrate the forms of Welsh which learners and children are most likely to come across when reading. Place of publication will be recorded in the metadata and a broad sample of written material will be sought overall to ensure that no one publisher dominates. The medium ‘Magazines’ includes the genre ‘papurau bro’, comprising of 60 local community Welsh-language newspapers/newsletters covering most areas of Wales which will capture a range of dialects and levels of proficiency.

E-language

It is only fairly recently that General, ‘National’ corpora have sought to include e-language data. According to Knight et al. (2014: 30) ‘while it is widely acknowledged that we live and communicate in a ubiquitous, digital world, the ways in which we actually do this, across multiple resources, remains an underexplored area of research in corpus linguistics as there is a lack of appropriate resources in existence to enable us to do this’.

CorCenCC will reflect modern usage of language and the use of electronic mediums to facilitate communication in the Welsh language context.

However, while it appears that e-language is all pervasive in English, this is not the case for Welsh. For example, there was a provisional plan to include language data from discussion boards in CorCenCC, but only four boards and forums conducted through the medium of Welsh were identified. Of those four, three had been dormant for many years. Clearly, care needs to be taken when deciding which elements should be included. Including only one community of people to represent an e-language type appears ill advised, thus discussion boards will not be included in the CorCenCC corpus.

Research by Beaufort Research (2013: 28) suggests that there is further disparity between the number of e-mails and SMS messages communicated in Welsh compared to English, even among Welsh speakers. 42% of Welsh speaking respondents to the Beaufort Research survey had sent a Welsh language e-mail in the previous month, with 72% having sent at least one in English. Furthermore, only 44% of respondents had sent a Welsh language SMS in the previous month. While this supports the inclusion of these e-language types in the corpus, it also suggests that the number of allocated words for these e-language types should be proportionally less than had the sampling frame been designed for an English corpus.

Each e-language type is categorised in respect of its general topic, purpose or content. Blogs and websites will be split into six categories, each containing 100,000 words, as illustrated in Figure 2.
<table>
<thead>
<tr>
<th>Code</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>News, Media and Current Affairs, Politics, Business and Finance, Weather and the Environment, Online Shopping</td>
</tr>
<tr>
<td>B</td>
<td>Religion, Language, Culture, Literature and the Arts, Teaching, Academia and Education</td>
</tr>
<tr>
<td>C</td>
<td>Technology, Computers and Gaming, Fashion and Beauty, Hobbies and Pastimes, Travel, Cookery</td>
</tr>
<tr>
<td>D</td>
<td>Music, Sport, Gigs and Events</td>
</tr>
<tr>
<td>E</td>
<td>Celebrity news and gossip, TV and Film, Humour</td>
</tr>
<tr>
<td>F</td>
<td>Parenting and Family Life, Health and Wellbeing, Personal and Daily Life</td>
</tr>
</tbody>
</table>

**Figure 2:** CorCenCC blog and website topics.

This thematic framework is broadly based on the one devised by the team constructing CANELC (Cambridge and Nottingham eLanguage Corpus) (see Knight et al. 2014), but with some important modifications to ensure that it is appropriate for the Welsh language context. For instance, religion was not included as a category in the CANELC framework. However, brief research conducted at the start of the CorCenCC project identified that a number of blogs were dedicated to discussing religious topics. This suggests that religion is an important domain in Welsh, and that it should therefore be included in the CorCenCC thematic framework. Of course, this framework is not appropriate for more private e-language types such as SMS and e-mail, which have been split according to the context in which these messages were sent, namely either business or personal.

In this presentation, the peculiarities of e-language use in the Welsh language context are explored, in respect of type and topic, both of which will be discussed in more detail.

**References**


\[1\] CorCenCC is funded by the Economic and Social Research Council (ESRC) and the Arts and Humanities Research Council (AHRC). Ref ES/M011348/1.
Expanding the coverage of a computational model for an endangered language with a derivational component – the case of Plains Cree

Antti Arppe, Katherine Schmirler (University of Alberta, Canada), Arok Wolvengrey (First Nations University of Canada, Canada), Miikka Silfverberg and Mans Hulden (University of Colorado, USA)

In the case of many endangered languages, dictionaries and lexical databases, if any exist at all, are typically substantially smaller in the extent of their lexical content – often in the range of 5-20 thousand lexical entries – in comparison to such resources for majority languages (e.g. English, Swedish, Finnish), which will usually contain at least 50-100 thousand lexical entries if not substantially more. Take for example Plains Cree (crk, Algonquian: Western Canada and United States), which among the endangered languages of the world is exceptionally fortunate in having four contemporary dictionaries, namely The Student's Dictionary of Plains Cree (Wolfart and Ahenakew, 1998, with 5,000 lexical entries), the Maskwacis Cree Dictionary (Miyo Wahkohtowin Education, with 8,986 lexical entries, nêhiyâwêwin: itwêwina / Cree: Words (Wolvengrey, 2001, with 16,453 lexical entries, cf. altlab.ualberta.ca/itwewina), and the Alberta Elders’ Cree Dictionary (Waugh et al., 2002, with 23,117 lexical entries). Now contrast these with the Oxford English Dictionary (with more than 600,000 words, cf. www.oed.com), Svenska Akademiens ordlista över svenska språket (SAOL, 14th edition, Svenska Akademien 2015, with 126,000 lexical entries), or the Kielitoimiston sanakirja [Dictionary of the Finnish Language Board] (Grönroos et al. 2016, with over 100,000 lexical entries).

But is the core vocabulary of Plains Cree truly substantially smaller than that of majority languages? What should be noted is that the three larger Plains Cree dictionaries mentioned above overlap only to a limited though varying degree. Of the 8,986 lexical entries in the Maskwacis Cree Dictionary, a minority of 2,845 (31.6%) cannot be traced back to lexical entries in Wolvengrey (2001), or their inflected forms, using a morphological parser (Snoek et al. 2014; Harrigan et al. 2017) based on the lexical content of Wolvengrey (2001). In contrast, as many as 15,623 (67.5%) of the 23,117 lexical entries in the Alberta Elders’ Cree Dictionary cannot be computationally derived as inflected forms of the lexical entries of Wolvengrey (2001). Based on our preliminary scrutiny of the non-overlapping lexical entries among these three dictionary resources, some are due to inconsistencies in orthography, and some represent genuine dialectal and areal differences in vocabulary. However, a large proportion of the non-overlapping entries can be considered part of the shared, core vocabulary of Plains Cree, and thus we estimate the overall vocabulary for Plains Cree, as manifested in these three dictionaries, to exceed 30,000 lexical entries. But might this start approaching the upper bound of contemporary Plains Cree vocabulary? Nowadays, in the case of majority languages, we are able to exploit ever increasing corpora to extract new vocabulary, but in the case of Plains Cree the corpus of the major known works excluding Bible translations and other religious texts, namely the texts compiled and edited by H. C. Wolfart and Freda Ahenakew as well as by Leonard Bloomfield, add up to only some 156,483 word tokens representing 18,605 word types, of which only a small fraction (364 word types) are currently not analyzable as Plains Cree word types. However, Plains Cree lexemes are almost entirely composed of native Cree morphemes, instead of loans from other Indigenous and majority languages, and moreover the derivation of Plains Cree words by concatenating these native morphemes continues to be a fairly regular and productive
morphological process, with the meaning of the resultant derived words being quite transparent, in contemporary Plains Cree (and generally for related Algonquian languages). Thus, we can use a computational model of derivational word/stem formation in Plains Cree to substantially expand our potential vocabulary coverage, though the practical usefulness of such a derivational model depends on how well we can incorporate information on what derivational morpheme combinations are the most likely ones, and consequently being able to rank the derivational decompositions of words as to their plausibility (Arppe et al. 2016).

Plains Cree stem derivation involves three subclasses of morphemes, initial or root morphemes, medial morphemes, and final morphemes. A stem usually involves an initial root morpheme; roots are generally not restricted to use in certain stem classes and so may often occur in nouns, verbs, and particles, though some roots are more restricted. Roots are followed by optional medial suffixes, which often occur with more concrete meanings and, like roots, are often not restricted to particular stem classes; for example, -âpisk(w) ‘metal’ can be used in nouns, as in pîwapisk ‘piece of metal’ and in verbs, as in kipâpiskaham ‘he closes it with metal, locks it’. Final suffixes are more restricted and are used to determine the class of a stem and, in verbs, the subclass of the verb: transitive, or intransitive, number of animate or inanimate arguments. The concatenation of a root, an optional medial, and a final morpheme constitutes a primary stem. This primary stem may then undergo further derivation, followed by an optional medial and a final morpheme, creating a secondary stem, which may also undergo further derivation (Wolfart, 1973, 1996).

While the computational modelling of inflectional morphology has resulted in a set of morphophonemic rules that can often be applied to Cree derivation, there are some other changes that are relevant to the modelling of derivational morphology. For example, -i is inserted between two consonants and word-final C_w becomes only C. An understanding of historical environments is also needed; for example, t may be palatalized to either c or s before i, or may remain unchanged. The conditioning factors have been obscured by sound change: t may be a reflex of either *θ or *t, and i may be a reflex of either *j or *e, which dictate the now-unpredictable alternations (Wolfart, 1996).

The lexical database underlying Wolvengrey (2001) presents an exhaustive derivational decomposition for some 10,363 verbs, consisting of combinations of 1,784 unique initial-like morphemic elements, 308 medial-like elements, and 457 final-like elements. This information together with the above morphophonemic rules can be used to create a general computational model as a finite-state transducer (Lindén et al. 2011) of how the derivational morphemes can be concatenated to form contemporary Plains Cree verb stems. For instance, in (1) we can see the possible derivational decompositional analyses for the stem acâhkosiwi- ‘it is a star; s/he is a star’. However, the analyses are unweighted, with equal status, not allowing us to know that /atâhkw-/is/-iwi/ is the most likely analysis. Fortunately, we can use the pairings of stems and their derivational decompositions provided in Wolvengrey (2001) as training data in order to weight all the morpheme sequences in the finite-state transducer as to their likelihood of occurrence (Mohri 1997; Pirinen 2014). As a result, this model can be used to provide all the possible derivational analyses of verb stems, which are ranked as to the overall likelihood of their constituent morpheme sequences, as well as generate the resultant stem form for any given allowable derivational morpheme sequence (2). Therefore, the correct derivational analysis also receives the best ranking with the weighted model in (2).
In this paper, we have explored how this weighted computational model can be used as an exploratory tool in the derivational analysis of previously undocumented Plains Cree verb stems. We are currently quantitatively evaluating with the non-overlapping forms from the three major Plains Cree dictionaries how well the derivational computational model fares in generally ranking the most plausible decompositional analyses. Such a derivational computational model could eventually be incorporated as a “guesser” within a general morphological analyzer to provide possible analyses for potentially grammatical words for which the full stem is lacking from the lexicon, which could substantially extend the coverage of such analyzers. Therefore, we hope to provide a case example of how the lack of extensive collections of lexicalized forms that are available for majority languages through extensive lists of lexicalized word forms in dictionaries as well as within large corpora can, in the case of less-resourced endangered languages that primarily use native morphemes for new word formation, such as Plains Cree, be compensated by a smaller lexical database, when that resource is comprehensively enriched with detailed derivational information, which allows for both generally specifying the computational derivational model as well as weighting its analyses based on the co-occurrence likelihood of derivational morphemes.

References


Continued gender inequality and gendered representations in the media, broadly construed, remain of concern because of the dialectic relationship between language and society. One source of gender cues is fiction written for and consumed by children. The characters encountered in the pages of a popular book constitute the stuff of identity building and may become role models for thousands of young and impressionable readers. Dialogue is “particularly important in fiction” (Sunderland 2011: 68) for the development of these characters, and may be even more so, and more common, in children’s fiction than in that for adults (Hunt 1991). How fictional characters’ utterances are described reveals authors’ assumptions about the ways in which females and males should or do speak in real life. The fact that reading acts as a source of socialisation for child readers means that patterns in the representation of male and female characters and the way they speak in children’s fiction can have significant consequences. This paper explores the expression of gender in verbs of speech (such as yell, shout, giggle or cry) in selected children’s fiction, focussing on a range of commercially successful series published in English in the last 70 years. The books are analysed using a blend of Corpus Linguistics and Critical Discourse Analysis.

The corpus consists of texts from a range of fictional series for children, including Enid Blyton’s Famous Five (published from 1942 to 1963), The Chronicles of Narnia by C.S. Lewis (1950 – 1956), Francine Pascal’s Sweet Valley series (1984-1998) and J.K. Rowling’s Harry Potter books (2001 to 2011). For both the Narnia and Potter series, the first, last and middle books in each series were chosen as these were the texts in which all the central child characters appeared prominently in the plot. One book from each of the Sweet Valley series (Sweet Valley Twins, SV High etc.) was chosen randomly for analysis. Each series was analysed separately to identify its unique patterns and then combined to form a corpus of nearly three million words which was analysed for general tendencies.

I used a variety of corpus software to identify trends in the representation of vocalisation in terms of the gender of the human characters in the stories. A basic frequency list allowed me to identify and rank the verbs, while ProtAnt 1.2.0 and AntConc’s 3.4.3 Concordance Plot revealed idiosyncratic frequency of usage in particular books. Sorting concordance lines enabled me to identify and quantify which verbs were typically attributed to female characters, and which to males, by grouping together tokens patterning with gendered pronouns, and character names. Sketch Engine provided Word Sketches which reveal the verbal behaviour of particular characters.

Motschenbacher (2009: 3) explores how nouns may be gendered in various ways, categorising these as Lexical Gender, Social Gender and Referential Gender, with a particular emphasis on body parts. Lexical Gender refers to personal nouns which are marked semantically as [male] or [female], such as uncle, grandmother, ovary; Social Gender encompasses “entrenched social stereotypes that tie certain role scripts to women and men” (op cit.: 4), such as eyelash or muscle, which,
despite not being physically restricted to one sex or the other, in practice tend to be associated with either the female or the male body, and Referential Gender is determined by the sex of the person who is being referred to. One could refer to a beard or muscles that belong to a woman, just as one could refer to these physical features on a man’s body; indeed this kind of gender could be used subversively to counter the expected references, as Motschenbacher points out, such as the use of sister amongst gay men. This raises the question of whether these types of gender, and especially social gender, which highlights entrenched social stereotypes which tend to be associated with either females or males, are found in the use of verbs as well. This paper begins to answer that question by exploring the gendering of verbs of speech, or reporting verbs, as used in fiction, and in children’s fiction in particular.

The results indicate clear polarisation between male and female characters, even in the more modern series of books. The similarities between the sub-corpora typically group in terms of the target audience, with the books for older readers or those written for the American market being quantitatively different, and therefore, I argue, supporting different worldviews, from the works for younger children by British authors, for instance. The SV books, for example, were generally more sexualised than the rest of the corpus, particularly the SV High and SV University titles. However, the relatively recent Harry Potter books are anachronistic both in terms of the fictional world they reflect (which is missing many modern features of the current real world, such as computers or mobile phones, while being replete with gargoyles and steam trains) and with respect to the traditional roles inhabited by female characters.

In terms of Caldas-Coulthard’s (1994) taxonomy of verbs of speech, metapropositional verbs (which could be considered speech acts, e.g. explain, agree, grumble, complain) and descriptive verbs (which include prosodic or paralinguistic information about the vocalisation, e.g. cry, shout, whisper, sigh) are the most frequent. The metapositional verbs frequently invoke power, or power relations, while the descriptive verbs tend to reflect emotion. In the corpus they both show strong gendered trends. For example, boasted, guffawed and bellowed are almost exclusively used by male characters, and giggled and chattered by females.

Amongst the other reporting verbs, there is also clear clustering into gendered discourse prosodies, such as the following:

<table>
<thead>
<tr>
<th>+ volume</th>
<th>Female characters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male characters</td>
<td>Female characters</td>
</tr>
<tr>
<td>bellow</td>
<td>screeched</td>
</tr>
<tr>
<td>roar</td>
<td>screamed</td>
</tr>
<tr>
<td>exploded</td>
<td>shrieked</td>
</tr>
</tbody>
</table>

In addition, pairs of speech acts are polarised in terms of gender, such as begged (68% female) versus ordered (77% male). Overall, the verbs linked to male characters are associated with loudness and power, while the ‘feminine’ verbs tend to feature trivial emotions or weakness, and high pitch. This shows clear gendering in terms of social meanings and expressive value, with the ‘masculine’ verbs being linked to positively valued notions of strength and those used by female characters seen as less important and weaker.
The current research does not distinguish between simple speaker attribution and more complex structures such as suspended quotations (Mahlberg, Smith and Preston, 2013), which are frequently associated with descriptions of the body (ibid.). In Hunt (2011), I analysed the gendered uses of body parts in children's fiction, and looking specifically at suspended quotations would be a fruitful blend of these two projects.

The children’s literature in English represented in this corpus reflects ‘traditional’ views of gender-appropriate ways of speaking, and reinforces them, supporting rather than contesting traditional gender roles. In the ideological context of the Western society in which the books were published, it is problematic that child readers are exposed to such polarised gender roles, which are rendered commonsensical by their reiteration in the stories. Harnessing the tools of corpus linguistics to analyse the books reveals the ideological discourses supported by the patterns of representation, in this case the gendered nature of reporting verbs, and allows for contestation and resistance by both the adults and children who read them.

References


Corpus-based resources for L1 and L2 teaching of Czech

Lucie Chlumská (Charles University, Czech Republic), Anna Čermáková (Charles University, Czech Republic) and Pavlína Vališová (Masaryk University, Czech Republic)

Introduction

Applications of corpus linguistics to language teaching began as early as the late eighties and nineties (e.g. Higgins 1988, Stevens 1991, J. Flowerdew 1993) and to this day, considerable amount of work has been done in this area, mostly in the second language acquisition and in teaching language for specific purposes. It is not surprising that it is English that has been in the centre of attention: as second language (ESL), as lingua franca (ELF), or as language for specific purposes (ESP). Currently, the learner corpora research is one of the most progressive areas of corpus linguistics (c.f. e.g. the ICLE\(^1\) project launched by Sylviane Granger and her colleagues at the University of Louvain).

However, the research on corpora in language teaching seems to have rather neglected the idea of using corpora and corpus-based resources in L1 teaching, especially at primary and secondary school levels. To our knowledge, the CLLIP project (Sealey & Thompson 2004) employing a corpus-based approach to teach 8-10 year-olds and the web-based Englicious project\(^2\) (developed by the team of Survey of English Usage, UCL) are one of the rare exceptions in the past few years.

This paper aims to introduce our joint effort to a) provide a basis for a theoretical framework to using corpus-based teaching materials for Czech, both as L1 and L2, incl. the identification of specific needs of L1 and L2 learners, and b) to present sample activities of real-life application of corpus-based exercises at the secondary school level (L1) and in the foreign learners’ classroom (L2).

Teaching Czech as L1 and L2

Czech presents, in comparison to English, a rather specific language situation: it is a ‘small’ language, which lacks some of the resources taken for granted in teaching e.g. English; there is no contemporary general dictionary, no school dictionaries, and the traditional mother tongue teaching curriculum in schools is heavily relying on teaching grammar and orthography rules and favouring literature lessons at the expense of language. According to the PISA cross-national assessment, Czech pupils are consistently below the OECD average in reading skills and text comprehension.

This urgent need to improve and change the mother tongue teaching in schools has been now acknowledged and an extensive research team from The

\(^1\) https://www.uclouvain.be/en-cecl-icle.html

\(^2\) http://www.englicious.org/
Faculty of Education\(^3\), Charles University, is currently mapping the situation in primary schools in order to suggest changes. These changes should also involve the development of new, corpus-based, resources.

Teaching Czech as a foreign language has also been rather traditional and conservative. Due to the fact that Czech is a highly inflected language, textbook exercises mostly focus on grammar drills and use invented model sentences and texts (very few textbooks use authentic examples). Learners of Czech as L2 tend to be of different mother tongues and it is common to have a very heterogeneous classroom – that is why teaching resources are not usually developed for a specific target group of non-native speakers of Czech.

One of our long-term objectives therefore is to make use of extensive corpus resources that are available for Czech, which include several national corpora, large parallel corpus and various specialised corpora, including e. g. a corpus of student writing. We aim to create teaching resources both for L1 and L2. Our efforts are currently being supported by various outreach activities aimed at both students and teachers; where – while raising awareness of corpus resources – the resources are at the same time being tested providing thus important user feedback. We have now established cooperation with the publishing house Fraus, a major Czech publisher of educational resources and developing further cooperation with the Faculty of Education within the above mentioned project. Currently the most visible output are corpus-based exercises incorporated into the Fraus textbooks and our platform [www.korpus.cz/proskoly](http://www.korpus.cz/proskoly) aimed at teachers (see below).

**Comparing the needs of L1 and L2 learners**

Since learners of Czech as L1 and L2 have different needs, it is necessary to first identify their specific nature and learning requirements. The following summary is based on our own hands-on experience.

In the L1 learning context, Czech is taught from the age of 6 to the secondary school graduate level of 18-19 year-olds. Corpus-based language exercises can be used at most levels (possibly excl. the youngest pupils who learn to read and write), e.g. our pilot project with 10-12 year-old students demonstrated that they could successfully navigate through the corpus interface and enjoyed working with real data (e.g. comparing selected features of spoken and written language, discovering word meanings in their natural context, or engaging in activities with a parallel corpus including Harry Potter series). However, at the same time it pointed to a need for a less complex interface should corpora be used hands-on during classes.

On the other hand, L2 learners of Czech are mostly adults and they learn Czech for very different reasons: most often it is their need for "survival" in a foreign country while studying/working in the Czech Republic. The main goal for everyone is to successfully communicate with native speakers. They especially need to learn how to combine words, how to use them in context and what is appropriate in different

\(^3\) Project PRIMUS/HUM/19 entitled Didaktika českého jazyka v současném vzdělávacím kontextu (Czech Language Didactics in Current Educational Context) led by Stanislav Štěpánik, PhD.
communicative situations, since colloquial, spoken Czech is very different from the formal, written language (on the morphological, syntactic as well as lexical level).

What both groups have in common is the need to master Czech orthography rules which may often seem counterintuitive and full of exceptions. The possibility to teach difficult spelling through concordances has already successfully been tested e.g. on French (Tyne 2016).

**Types of corpus-based materials**

Corpora in teaching can be used both directly (“hands-on”, using computers/tablets/mobile phones) and indirectly (“paper” exercises based on corpus data). The first option seems to be at the moment less favoured by teachers (due to too complex corpus query interface and computer lab accessibility); however, based on our experience, students themselves enjoy working with some of our online corpus-based tools, e.g. SyD – a corpus-based tool comparing variants in spoken and written Czech. To support the second, more favoured, option, i.e. ready-made exercises, we have launched an online repository of ready-made L1 and L2 corpus-based exercises for teachers⁴.

In general, we have been working with the following typology of corpus-based exercises (based on Vališová 2016) with the aim to identify the most suitable resources for both L1 and L2. In addition to the direct approach (searching corpora), the corpus-based materials can be based on e.g. 1) observation – a selection of concordance lines with one word or a phrase, two words / phrases for comparison, 2) gap-filling – one or more words missing in the concordance lines, 3) mixed concordances – matching the corresponding left and right context, 4) identification – matching a node word with its collocates or assigning the right text type, 5) mix – combination of the above-mentioned types.

**Sample activities**

In the L1 environment, we have provided corpus-based exercises (mostly focusing on lexicology, word formation, or orthography) for Czech language textbooks for 7th-9th grade (13-15-year old students) as well as for secondary schools (examples will be provided in the presentation). Currently, we are testing the corpus-based approach on the secondary school level in direct cooperation with students themselves – presenting findings about language to their classmates and assessing the difficulty of exercises.

Our L2 activities are based on the corpus of written contemporary Czech and consist of authentic or partly edited sentences. Some of the exercises have already been successfully tested in an English speaking classroom, A2 level. The exercises focus on problematic words and their use (e.g. žít vs. bydlet – both translated as ‘to live’ into English). The testing showed that the ideal model is to introduce several sentences in concordances to observe the behaviour of the word (its meaning and

---

⁴ www.korpus.cz/proskoly
typical context) and then to present more examples in gap-filling exercises. Gap-filling exercises are more demanding for the teacher to prepare, as it is necessary in Czech to put into the brackets additional morphological information such as person, singular / plural, or tense for verbs. However, the great benefit is in practicing meaning and grammar at the same time.

References

The evaluative know-how of Chinese scholars: A contrastive corpus analysis of evaluative *it* patterns in research articles

Bingxin Wang and Naixing Wei (Beihang University, China)

1. Introduction

In recent years, evaluative *it* patterns have attracted broad attention in the literature. It is a prevalent language phenomenon in academic discourse (Biber et al., 1999; Hewings & Hewings, 2002; Peacock, 2011), and part of the defining code of research articles (RAs hereafter) (Williams, 2002). It allows writers to present their evaluation explicitly while giving them the appearance of generality and objectivity (Herriman, 2000) and is thus a powerful and effective construction for expressing attitudinal meanings.

Evaluative *it* patterns have been explored across populations, genres and disciplines (e.g., Hewings & Hewings, 2002; Peacock, 2011; Zhang, 2015). Yet, scant attention is paid to the characteristic patterns and their characteristic meanings in academic texts. Characteristic patterns (CPs hereafter) refer to the high-frequency patterns in a specific set of texts, while characteristic meanings (CMs hereafter) are the meanings expressed by the CPs. CMs generally reveal the typical meanings and functions of the texts, i.e., “what is often said” in the texts (Hunston, 2011). Francis et al. (1998) proposes 9 high-frequency evaluative *it* patterns in general English texts, including *it* v-link ADJ to, *it* v-link ADJ that, *it* v-link ADJ wh-, *it* v-link ADJ for n that, *it* v-link ADJ of n that, *it* v-link ADJ to n that, *it* v-link ADJ for n to, *it* v-link ADJ of n to and *it* v-link ADJ v-ing. However, whether these patterns are CPs in academic texts or not and what are the CMs of *it* patterns remain unanswered.

Besides, the majority of the existing literature focuses on the practices of native speakers, while how the patterns are used in non-native speaker writers’ academic texts, in Chinese scholars’ texts especially, and whether the usage of *it* patterns by Chinese scholars varies from that by Western scholars are still under-researched.

We thus set out to explore how Chinese scholars use evaluative *it* patterns to express attitudinal meanings and compare their practices with those of Western scholars. The following research questions are addressed:

1. What are the characteristic evaluative *it* patterns and their characteristic meanings in academic texts?
2. What similarities and differences exist in the use of characteristic *it* patterns by Chinese scholars and Western scholars in terms of frequency and characteristic meanings?

2. Method

The data for this study comes from Beijing Collections of Academic Research Essays (Beijing CARE). Beijing CARE consists of two comparable sub-corpora: the first one comprising 2,249 RAs produced by Chinese scholars (CHC), totaling 10,794,193 tokens, and the second one comprising 3,923 RAs by English native-speaker scholars (NSC), totaling 22,636,925 tokens. The RAs in the corpus were collected from English-
medium international leading journals in 23 fields in the period 2000 to 2014. Authors in CHC are based at a university in mainland, China, while those in NSC are affiliated with institutions of English inner circle countries, including United Kingdom, USA, Canada, New Zealand and Australia (Kachru, 2003). Two versions of Beijing CARE are provided: one with plain texts and one with POS-tagged texts.

We adopted the Pattern Grammar (Hunston & Francis 2000) approach in this study. The target patterns were \textit{it v-link ADJ to}, \textit{it v-link ADJ that}, \textit{it v-link ADJ wh-}, \textit{it v-link ADJ for n that}, \textit{it v-link ADJ of n that}, \textit{it v-link ADJ for n to}, \textit{it v-link ADJ of n to} and \textit{it v-link ADJ v-ing}. All variants of the linking verbs were included, including \textit{is}, \textit{was}, \textit{seems}, \textit{appears}, \textit{would be}, \textit{may be}, \textit{might be}, etc., as well as their corresponding negative forms. WordSmith Tools (v. 6.0, Scott, 2012) were first used to retrieve and count all occurrences of the \textit{it} patterns in the tagged version of Beijing CARE. Tagged sequences (e.g., \textit{it\textunderscore pph1 is\textunderscore vbz \_\_\_\_\_ to\_\_\_\_ to}) were used to concordance the lexical sequences (e.g., \textit{it is possible to}) of specific patterns (e.g., \textit{it is ADJ to}). Next was a search for the high-frequency adjectives in the patterns through the plain text corpus. The cut-off frequency was set at 10 per million words. Based on Francis et al. (1998), we categorized the high-frequency adjectives into meaning groups according to their semantic features. Then the evaluative meanings of \textit{it} patterns across sub-corpora were analyzed and compared. As the two sub-corpora are of different size, normed figures (calculated per million words) are used to report the frequency data. Statistical significance (p<0.05) was tested with the loglikelihood-ratio test.

3. Results

3.1 Distribution of evaluative \textit{it} patterns

Table 1 reports the distribution of the evaluative \textit{it} patterns in the corpus.

<table>
<thead>
<tr>
<th>\textit{it} patterns</th>
<th>CHC Raw</th>
<th>CHC Normed</th>
<th>NSC Raw</th>
<th>NSC Normed</th>
<th>Loglikelihood ratio</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>\textit{it v-link ADJ to}</td>
<td>3,575</td>
<td>331</td>
<td>8,834</td>
<td>390</td>
<td>-70.06</td>
<td>0.000</td>
</tr>
<tr>
<td>\textit{it v-link ADJ that}</td>
<td>1,741</td>
<td>161</td>
<td>5,004</td>
<td>221</td>
<td>-134.65</td>
<td>0.000</td>
</tr>
<tr>
<td>\textit{it v-link ADJ for n to}</td>
<td>296</td>
<td>26</td>
<td>454</td>
<td>19</td>
<td>16.92</td>
<td>0.000</td>
</tr>
<tr>
<td>\textit{it v-link ADJ wh-}</td>
<td>139</td>
<td>13</td>
<td>476</td>
<td>21</td>
<td>-28.13</td>
<td>0.000</td>
</tr>
<tr>
<td>\textit{it v-link ADJ to n}</td>
<td>3</td>
<td>0</td>
<td>17</td>
<td>1</td>
<td>-3.14</td>
<td>0.076</td>
</tr>
<tr>
<td>\textit{it v-link ADJ for n that}</td>
<td>1</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>-1.20</td>
<td>0.273</td>
</tr>
<tr>
<td>\textit{it v-link ADJ of n that}</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>-4.69</td>
<td>0.030</td>
</tr>
<tr>
<td>\textit{it v-link ADJ of n to}</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.00</td>
<td>1.000</td>
</tr>
<tr>
<td>\textit{it v-link ADJ v-ing}</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.00</td>
<td>1.000</td>
</tr>
<tr>
<td>\textit{Total}</td>
<td>5,755</td>
<td>533</td>
<td>14,797</td>
<td>654</td>
<td>-177.07</td>
<td>0.000</td>
</tr>
</tbody>
</table>

As can be seen in Table 2, there are four high-frequency evaluative \textit{it} patterns in both
sets of texts: **it v-link ADJ to**, **it v-link ADJ that**, **it v-link ADJ for n to** and **it v-link ADJ wh-**. The distribution trend of evaluative *it* patterns in academic texts found in this study differs greatly from that in general English texts, where, as Francis et al. (1998) notes, *9 it* patterns figure prominently as the major evaluation devices. Yet, our study suggests that there are only four dominant *it* patterns in academic texts. These four patterns are the CPs in academic texts.

Striking differences are observed between the two sub-corpora. The occurrence of *it* patterns in NSC significantly outnumbers that in CHC. Also, in CHC, the frequencies of **it v-link ADJ to**, **it v-link ADJ that** and **it v-link ADJ wh-** are significantly lower than those in NSC. It is not surprising to find this trend, given that the overall frequency of evaluative *it* patterns in CHC is much lower. Yet, interestingly, CHC displays a significantly higher frequency of **it v-link ADJ for n to**.

### 3.2 CMs of **it v-link ADJ to** and **it v-link ADJ that**

Figure 1 presents the meaning groups associated with **it v-link ADJ to** pattern across sub-corpora. The two groups of scholars employ the pattern to convey similar evaluative meanings, that is, to assess the importance, likelihood, difficulty, desirability and usefulness of acts or procedures. These attitudinal meanings, constituting the common traits of both data sets, are CMs in academic discourse.

![Figure 1. The meaning groups of **it v-link ADJ to** pattern](image)

Besides, Chinese scholars differ from Western scholars as regards the semantic preferences of this pattern. Western scholars evaluate the importance, likelihood and usefulness with significantly greater frequency. Among these evaluative meanings, difference in the likelihood group is found to be most statistically significant (LL=-417.20, p<0.001); Western scholars are on average three times more likely to indicate the likelihood of acts. Chinese scholars, by contrast, express difficulty more frequently (LL=100.53, p<0.001).

Figure 2 gives the meaning groups of **it v-link ADJ that** across sub-corpora. Both Chinese scholars and Western scholars use this pattern to evaluate the certainty, likelihood, importance and expectation of propositions. Certainty and likelihood are the two dominant meaning groups in both sets of texts. The preference for the certainty and likelihood meanings is a commonality of academic writing.
While the two sub-corpora contain similarities, there are considerable variations in the distribution pattern of the meaning groups. In contrast with Western scholars, Chinese scholars use this pattern less frequently to express meanings of likelihood, importance and expectation, but more often to indicate the certainty of their propositions. All differences were found to be statistically significant, among which difference in the likelihood group is the most remarkable one (LL=-311.42, p<0.001).

4. Discussion

Our results reveal a number of features of evaluative *it* patterns in RAs produced by Chinese scholars and their Western counterparts. The commonly used CPs and CMs in the two sets of data are indicative of the shared ideology and academic conventions of the academic community. The differences between the two groups of writers, however, can be possibly attributed to a series of factors. First, the deficiency in pragmatic competence and L2 knowledge may hinder Chinese scholars from expressing attitudinal meanings appropriately. Second, culturally preferred epistemological beliefs and rhetorical conventions may be an influencing factor of the rhetorical choices of the writers.

References


“Please please don’t buy this game like I did. I feel terrible and wish I could return it!”: A corpus-based study of professional and consumer reviews of video games
Andrew Kehoe and Matt Gee (Birmingham City University, UK)

Background

This paper is a corpus-based comparison of professional and consumer reviews of video games. Professional video games journalism is a well-established field, with the first dedicated publications launched in 1981 (Computer and Video Games in the UK, followed by Electronic Games in the US). Over the next 30 years, such publications were highly influential in the success of particular video games, often leading to accusations of conflicts of interest, with the publishers of the games under review also paying substantial amounts of money to advertise their products in the same video games magazines.

The rapid growth of the web since the early 2000s has seen a decline in all traditional print media, and games magazines are no exception. Such publications have now been largely replaced by professional video games review sites such as IGN.com, established in 1996 as the Imagine Games Network. However, the influence of professional review sites such as IGN appears to be waning, with a recent report by the Entertainment Software Association suggesting that only 3% of consumers rely on professional reviews as the most important factor when making a decision on which game to purchase (ESA 2015).

The decline of professional games reviews does not reflect the state of the video games industry in general. It has been estimated (Newzoo 2016) that the industry generated almost $100 billion of revenue worldwide during 2016. At the same time, we have seen a growth in consumer reviews of products online, on ecommerce sites such as Amazon but also on dedicated review sites in particular industries. For video games, the predominant site is Metacritic.com, which aggregates professional reviews from various sources but also provides a platform for gamers to give their own reviews.

Anecdotally, there is seemingly growing discord between professional reviews and consumer opinions on the same games. As far as we are aware, though, there have been no previous linguistic studies comparing professional and consumer reviews of video games. There has been some research on professional and consumer reviews of films (de Jong & Burgers 2013). This work was a genre analysis of 72 reviews, relying on manual coding of descriptive and evaluative moves (e.g. ‘placing the movie in context’ and ‘recommending the movie’ respectively). The authors found that professional reviewers tend to give more practical information about the film whereas consumer critics use more first-person, evaluative language. However, the scope of the study was necessarily limited by their manual approach. In our research, we adopt a corpus-driven approach, focussing on lexical differences in a large corpus of professional and consumer reviews.
The Corpus

Our corpus contains online reviews by professional critics and reviews written by consumers on the Metacritic website. We included only games on the PlayStation 4 and XBox One platforms which had been reviewed by all of the four top ranked (according to Alexa web rankings) online games magazines: IGN, GameSpot, DigitalSpy, and GamesRadar. A maximum of 100 user reviews were downloaded for each game on each platform, based on the default sort order used by Metacritic (a rating of the ‘most helpful’ reviews).

The resulting corpus covers 150 games, with the professional reviews sub-corpus containing 707,336 tokens (from 593 reviews) and the consumer review sub-corpus containing 2,184,748 tokens (from 14,771 reviews).

Key Words Analysis

Table 1 shows a Key Word analysis (Scott 1997) of the professional reviews compared against the consumer reviews. We see that professional reviewers tend to be preoccupied with giving factual information about the game, whether this be the developer and publisher or the type of game (single-player, open-world, first-person). The consumer reviews, on the other hand, tend to contain more evaluative language: good, fun, great, amazing, bad, love, boring, awesome, best. This is similar to the findings of de Jong & Burgers (2013) in their analysis of film reviews but on a much larger scale and without the need for time-consuming manual coding.

Table 1 Key words from the Professional and Consumer reviews

<table>
<thead>
<tr>
<th>Professional Reviews</th>
<th>Consumer Reviews</th>
</tr>
</thead>
<tbody>
<tr>
<td>game’s moments three particularly enemies sense single-player two enemy new attacks four abilities open-world nearby ability occasionally events rise players</td>
<td>together place narrative foes action power distinct attack instance meanwhile in-game battles fire five first-person process individual specific course means</td>
</tr>
</tbody>
</table>
In our key word analysis we see consumers make greater reference to specific elements of the game: **graphics, story, gameplay, bugs.** These findings echo the ESA report (ESA 2015) which found that the largest proportion of consumers (22%) rank “interesting story/premise” as the most important factor in making a purchase decision. Our analysis also chime with an online news article on the ESA report (Breitbart 2015) which states that professional reviews “have long been criticised for being overly politicised and for focusing on irrelevant detail at the expense of conveying information about frame rates, gameplay and other factors gamers say they care about”.

We also see stylistic differences between the two sub-corpora, with evidence of a more formal writing style in the professional reviews: particularly, sense, occasionally, instance, distinct, meanwhile. Consumer reviews tend to be less formal, as seen in the following example:

_Wont be buying games on release day any more SADLY !!!!! I wanted to love this game & give it a 10 but IT's A BIG FAIL !!!_

However, as we go on to demonstrate it would be inaccurate to generalise and claim that all consumer reviews are poorly written and lacking in depth. Using examples from our corpus, such as the following, we illustrate that many of the rhetorical features more commonly associated with professional reviews are present in consumer reviews too:

_In closing, the LOTR fan will love this game, and any fans of the Assassin’s Creed or Batman series could definitely get into it._

Through a subsequent analysis of key n-grams we explore in depth further differences between professional and consumer reviews. For example, the 3-gram _buy this game_ is significantly more frequent in consumer reviews, with both negative (as in the title of the paper) and positive examples (“Go and buy this game now, it’s awesome!”). Such specific recommendations tend to be absent in professional reviews.

**Differences between Professional Review Sites**

In the final part of the paper we explore linguistic differences between the four professional review sites in our corpus (Table 2). We find evidence of a more informal first-person style in IGN reviews (_I, me, my, I’ve, I’d, myself_), as well as a use of evaluative language which is more commonly found in consumer reviews (_fun, great, interesting_). This is in sharp contrast to previous research on professional film reviews (de Jong & Burgers 2015).

As we illustrate, this seems to be the result of an increasing trend on the IGN website for reviews in video format with a text transcript below or, more accurately, for reviews which are written to be read out loud. The language of such reviews remains the conventional language of professional reviews but modified slightly to make it more personal and more suitable to be shared in video form, e.g.

_Assassin’s Creed Chronicles: China is a decent attempt at converting the Assassin’s Creed experience into a side-scrolling form-factor, but it lacks heart._
The pieces are all there, but they are patched together in a dry, passionless way that left me feeling emptier the longer I played it. [http://uk.ign.com/articles/2015/04/21/assassins-creed-chronicles-china-review](http://uk.ign.com/articles/2015/04/21/assassins-creed-chronicles-china-review)

Table 2 Key Words for each publication in the corpus.

<table>
<thead>
<tr>
<th>IGN</th>
<th>GameSpot</th>
<th>DigitalSpy</th>
<th>GamesRadar</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>you</td>
<td>we</td>
<td>you’ll</td>
</tr>
<tr>
<td>me</td>
<td>your</td>
<td>players</td>
<td>reviewed</td>
</tr>
<tr>
<td>my</td>
<td>game’s</td>
<td>game</td>
<td>will</td>
</tr>
<tr>
<td>I’ve</td>
<td>these</td>
<td>our</td>
<td>you’re</td>
</tr>
<tr>
<td>I’d</td>
<td>her</td>
<td>we’d</td>
<td>starts</td>
</tr>
<tr>
<td>version</td>
<td>may</td>
<td>although</td>
<td>game</td>
</tr>
<tr>
<td>playstation</td>
<td>must</td>
<td>games</td>
<td>yes</td>
</tr>
<tr>
<td>myself</td>
<td>such</td>
<td>we’ve</td>
<td>that’s</td>
</tr>
<tr>
<td>fun</td>
<td>fight</td>
<td>it’s</td>
<td>via</td>
</tr>
<tr>
<td>like</td>
<td>moments</td>
<td>we’re</td>
<td>theme</td>
</tr>
<tr>
<td>great</td>
<td>nearby</td>
<td>fans</td>
<td></td>
</tr>
<tr>
<td>found</td>
<td>process</td>
<td>really</td>
<td>split</td>
</tr>
<tr>
<td>didn’t</td>
<td>he</td>
<td>something</td>
<td>up</td>
</tr>
<tr>
<td>was</td>
<td>how</td>
<td>will</td>
<td>kicking</td>
</tr>
<tr>
<td>felt</td>
<td>upon</td>
<td>visuals</td>
<td>incredibly</td>
</tr>
<tr>
<td>interesting</td>
<td>lead</td>
<td>colourful</td>
<td>bits</td>
</tr>
<tr>
<td>it</td>
<td>several</td>
<td>you’ll</td>
<td>know</td>
</tr>
<tr>
<td>had</td>
<td>many</td>
<td>has</td>
<td>real</td>
</tr>
<tr>
<td>versions</td>
<td>I</td>
<td>users</td>
<td>start</td>
</tr>
<tr>
<td>mostly</td>
<td>ordinary</td>
<td>colour</td>
<td>damn</td>
</tr>
</tbody>
</table>

**Conclusion**

In this paper we present novel research on the application of corpus linguistic techniques to a ‘real-world’ problem of growing importance: the mining of consumer intelligence data. With reference to the domain of video games, we demonstrate how the relevance of professional reviews in the buying process has decreased while consumer reviews have grown in importance. We show that the latter tend to be less formal and more personal in style, and tend to have different preoccupations: graphics, gameplay and fun. We also show that some of the professional review sites that remain appear to be adapting to reflect the more immediate, personal style of consumer reviews.

**References**


**MI-score-based collocations in language learning research: A critical evaluation**

Dana Gablasova, Vaclav Brezina and Tony McEnery (Lancaster University, UK)

**Introduction and motivation**

Formulaic language has occupied a prominent role in the study of language learning and use for several decades (Wray, 2013). Recently an even more notable increase in interest in the topic has led to an ‘explosion of activity’ in the field (Wray, 2012, p.23). Language learning research (LLR) in both first and second language acquisition has focused on examining the links between formulaic units and fundamental cognitive processes in language learning and use, such as representation of and access to these units in mental lexicon (Wray 2002, 2012, 2013; Ellis et al, 2015). Collocation, a specific unit of formulaic language, holds a prominent position in LLR, having been used in a number of studies on formulaicity in L2 (Schmitt, 2012). The statistical measures for identifying collocations, association measures (AMs), in these studies are of paramount importance as they directly and significantly affect the findings of these studies and consequently the insights into language learning that they provide. One of the most prominent and frequently selected association measure in these studies is the Mutual Information score (MI-score), often referred to as a measure of collocational ‘strength’ (c.f. Hunston, 2002).

While MI-score has been a useful measure in LLR, there are also several issues related to its use (Gablasova et al. forthcoming 2017). First, the rationale behind the selection of MI-score in studies on formulaic development is not always fully transparent and systematic (González Fernández & Schmitt, 2015) and is often motivated by tradition rather than by specific aims of a given LLR study. Second, alternative measures are rarely considered and their relevance to LLR is not further examined (Gilquin & Gries, 2009). Finally, the application and interpretation of MI-score in LLR suggests that a fuller understanding of the mathematical and linguistic principles on which the measure is based is needed in LLR studies (e.g. an understanding of what type of collocations receive higher MI values and the reasons for this). This understanding would enable a better interpretation of collocational patterns found in L2 production.

**Research aims & methodology**

This study offers an empirical validation of MI-score based collocations for LLR research. In order to address the above issues, the paper seeks to achieve the following three objectives: i) to place MI-score in the context of other similar association measures and discuss the similarities and differences directly relevant to LLR; ii) to examine the effect of a specific corpus (and register/genre) on the MI values and discuss the implications of the differences in collocational strength measured by MI-score in these corpora; iii) propose general principles for selection of association measures in LLR. The study examines these questions using data
from several corpora and sub-corpora (the BNC and its written and spoken sub-corpora, CANCODE and the spoken component of BNC2014 – the newly developed corpus of British English) (see Table 1 for an overview). The whole BNC was included as it has traditionally been used in language learning collocational research as a reference corpus (e.g. Durrant & Schmitt, 2009; Granger & Bestgen, 2014). Five BNC sub-corpora were used to investigate the variation in collocational strength inside the BNC. CANCODE and BNC_SP were selected to strengthen the spoken component of the study by looking at more recent corpora of informal speech which are directly comparable with the informal sub-corporus (i.e. BNC-Demographic) from the BNC.

Table 1 Overview of (sub) corpora used

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Size</th>
<th>Representativeness</th>
</tr>
</thead>
<tbody>
<tr>
<td>British National Corpus (BNC)</td>
<td>98,560,118</td>
<td>Written and spoken (10M), diff. registers</td>
</tr>
<tr>
<td>BNC_A</td>
<td>15,778,043</td>
<td>Written, academic writing</td>
</tr>
<tr>
<td>BNC_N</td>
<td>9,412,245</td>
<td>Written, news</td>
</tr>
<tr>
<td>BNC_F</td>
<td>16,143,913</td>
<td>Written, fiction</td>
</tr>
<tr>
<td>BNC – Context governed (BNC_CG)</td>
<td>6,196,134</td>
<td>Spoken, formal</td>
</tr>
<tr>
<td>BNC – Demographic (BNC_D)</td>
<td>4,234,093</td>
<td>Spoken, informal</td>
</tr>
<tr>
<td>BNC – 2014 Spoken (BNC_SP)</td>
<td>4,789,185</td>
<td>Spoken, informal</td>
</tr>
<tr>
<td>CANCODE (CANC)</td>
<td>5,076,313</td>
<td>Spoken, informal</td>
</tr>
</tbody>
</table>

We selected three types of collocations representing a range of constructions that commonly appear in language learning collocational research (e.g. Siyanova & Schmitt, 2008; Durrant & Schmitt, 2009; Paquot & Granger, 2012; Granger & Bestgen, 2014, Ebeling & Hasselgård, 2015): verb + complementation (make + sure/decision/point), adjective + noun (human + beings/rights/nature) and adverb + adjective (vitaly/very/really + important). Using the selected (sub) corpora, we examined the strength of these collocations using MI-score and contrasted it with two other association measures (Log Dice and t-score) – an example of the results can be seen in Table 2 which illustrates the difference between three collocational measures across different sub(corpora). Special attention was paid to how collocational patterns (i.e. collocational strength) change according to a different genre/register or mode of communication (i.e. written vs spoken language).

Results

The following table provides a brief outline of the results obtained with one of the variables (make + complementation) that we investigated as an example.

Table 2 Make (lemma), [L0 R2]

<table>
<thead>
<tr>
<th></th>
<th>BNC</th>
<th>BNC_A</th>
<th>BNC_N</th>
<th>BNC_F</th>
<th>BNC_CG</th>
<th>BNC_D</th>
<th>BNC_SP</th>
<th>CANCODE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sure</td>
<td>6.80</td>
<td>7.09</td>
<td>7.26</td>
<td>5.78</td>
<td>6.90</td>
<td>6.64</td>
<td>6.26</td>
<td>6.92</td>
</tr>
<tr>
<td>decision</td>
<td>4.55</td>
<td>3.67</td>
<td>4.07</td>
<td>5.86</td>
<td>6.12</td>
<td>7.91</td>
<td>7.57</td>
<td>8.07</td>
</tr>
<tr>
<td>point</td>
<td>3.44</td>
<td>2.92</td>
<td>3.84</td>
<td>3.68</td>
<td>4.11</td>
<td>3.12</td>
<td>3.01</td>
<td>3.93</td>
</tr>
</tbody>
</table>
As can be seen from the table, there are differences between the strength of association between two words according to the measure used and according to the (sub)corpus in which the association was measured. While the difference between measures such as t-score and MI-score was expected, the difference between MI-score and Log Dice deserves further attention as both measures reward similar linguistic properties of collocations (e.g. exclusivity of association). The presentation will discuss the individual results including multiple variables in detail; due to space constraints these could not be fully included in the abstract.

Discussion

With respect to the three research aims, the results revealed:

i) A difference between the three AMs (MI-score, Log Dice and t-score) in identifying the strength of the relationship between words; in particular, the difference between MI-score and LogDice is interesting and has implications for the selection and interpretation of AMs in language learning research.

ii) The variation in the collocational strength between the BNC and its various subcorpora suggests that large aggregate data such as the whole BNC may hide different distributions of formulaicity across registers and genres as well as across the written/spoken divide.

iii) Following these findings, we propose general principles for the selection of AMs for language learning research. These include the need to understand 1) the mathematical reasoning behind the measure, 2) the scale on which it operates and 3) its practical effect (what combinations of words get highlighted and what gets hidden/downgraded).

References


Searching for the unsearchable: A study of English NP inversion
Amanda Patten (University of Birmingham, UK)

This paper outlines the methodological challenges of undertaking a corpus investigation of English NP inversions, such as (1).

(1) A particularly striking feature of the report is the growth in coverage in manual operations (ICE-GB)

Here, the initial indefinite noun phrase seems to be predicated of the postverbal NP (the logical subject); and the sentence is also specifying (identifying) in function. Thus, this construct is relevant both to work on full inversion, in which subject and complement are inverted (as in (2), involving a preverbal locative expression), and to accounts of specificational copular sentences, like (3), which contain preverbal definite NP subjects.

(2) In the radio car is Sir Geoffrey Johnson-Smith (ICE-GB)
(3) The lead actress in that movie is Ingrid Bergman (Mikkelsen 2005:1)

Indefinite NP BE NP sentences have held a marginalised position in data-oriented studies of inversion. The difficulty in reliably distinguishing which NP is subject and which complement has led some authors to exclude this sentence type from corpus analysis (e.g. Dorgeloh 1997); it also calls into question their representation in collections of inversion sentences drawn from opportunistic encounters in everyday reading (e.g. Birner 1996; Chen 2003). Certainly, in many studies, there is evidence of inconsistency in the criteria used to include or exclude individual NP BE NP tokens. Part of the problem seems to be that authors often discuss indefinite NP inversion without considering its relationship to specifications. Prado-Alonso (2011: 29) notes simply that these constructions have “sometimes been confused”.

However, there are many theoretical accounts of specificational sentences as instances of inversion (rather than as identity structures). Their challenge is to explain why inversion is almost always possible for sentences with definite NP predicates, but is often ungrammatical for sentences with indefinite NP predicates, shown in (4). These accounts very rarely engage with empirical studies of English inversion constructions.

(4) *A doctor is John (Mikkelsen 2005: 154)

In Author (2016), I claim that the acceptability of indefinite NP inversions depends both on the discourse status of the initial noun phrase – as being anchored to the prior text (see Mikkelsen 2005; Birner 1996) – and on the distinctiveness (or narrowness) of the description in characterising the logical subject. This is consistent with their treatment both as specificational copular sentences and as inversion constructions.

In this study, I use electronic corpus data to test this theory – a task which presents numerous methodological challenges. Searching for NP BE NP structures produces vast amounts of unwanted tokens which exhibit the canonical, non-inverted configuration. As a fully parsed corpus, the ICE-GB is designed to allow for the simple extraction of inversions, through either the INVOP function or INV/PRECS clause structure. However, upon investigation, I find that this subjective annotation is attached to relatively few viable instances of NP inversion in the corpus. This calls into question the frequencies and conclusions drawn from this approach (employed in Prado-Alonso 2011).
An alternative option is to specify the make-up of the initial NP in the form of a detailed search string (as employed by Kreyer 2006). However, since the focus of interest here is the complexity and variety of the initial description, such an approach has obvious limitations. In this paper, I discuss my attempt to achieve total accountability in relation to the corpus (McEnery and Hardie 2012). My approach combines a series of exploratory searches on syntactic structure with a careful manual sorting and selection process.

Upon obtaining a comprehensive data set from the ICE-GB, I focus on the formal and semantic characteristics of the initial indefinite noun phrase, and find that the construction shows a preference for modified NPs that provide a distinctive (if not inclusive) description of the logical subject. In some instances, this serves to override the principle of end weight, which is often called upon to explain inversion phenomena.

My data also point to the existence of subtypes of indefinite NP inversion which have construction-specific properties. Adopting the framework of construction grammar, I situate these constructs within a network of related constructions, focusing in particular on recognised points of difference between full inversion and specificationals, such as the subject status of the preverbal and postverbal NPs. My analysis clarifies the relationship between these constructions, having implications for both fields of study.

References

The semantic patterning of grammatical keywords in undergraduate History and PIR (Politics & International Relations) essays: a corpus-driven investigation
Karin Whiteside (University of Reading, UK)

The poster reports results from a PhD project completed at the University of Warwick which involved a comparative lexico-grammatical analysis of third-year student writing belonging to the Essay genre family (Nesi and Gardner, 2012) in two disciplines, History and PIR (Politics and International Relations), from two UK higher educational institutions. The project adopted a corpus-driven approach which was developed by Groom (2007) in his analysis of professional academic writing in Literature and History: a keyness analysis was used to identify statistically significant grammatical words, and the phraseological patterning around these was then qualitatively analysed and phraseologies categorised according to their semantic purposes. The concepts of ‘semantic sequence’ (Groom, 2007; Hunston, 2008) and ‘semantic motif’ (Groom, 2007), with their focus on ‘semantic similarity but formal variation’ (Hunston, 2008, p. 292) are crucial in this bottom-up phraseological method: they allow a wide-lens view of groupings of semantic functions and the range of ways in which these can be realised linguistically within texts.

In the project five grammatical keywords - of, and, that, as and this - were analysed across four sub-corpora each consisting of student writing from one of the two disciplines at one of the two institutions. It was found that there were more similarities than there were variations in the semantic patterning of grammatical keywords across the four disciplinary/institutional sub-corpora, and that these similarities could to a large extent be explained in terms of the shared features of student Humanities and Social Sciences writing (Durrant, 2017). The variations that occurred fell along disciplinary rather than institutional lines and can to an extent be explained by differences in the kinds of essay focus favoured by the two disciplines (e.g. focus on specific actors, contexts or phenomena in History vs. focus on general concepts in PIR) and can also be linked to Gosden’s (1993) model of subject role domains.

It is argued that Groom’s (2007) approach is a useful one to take in analysis of student writing as it uncovers lexico-grammatical features which occur very regularly within student texts and thus, from a pedagogical perspective, are of high value in terms of how much of the text they ‘operationalize’ (Bruce, 2011, p. 6).

References


A Comparable Corpus of Original and Translated Lithuanian:  
Design and Preliminary Findings  
Jurgita Vaičenonienė (Vytautas Magnus University, Lithuania)

In Corpus-Based Translation Studies, it is claimed that there are certain features common to all translations which make them different from original texts. Therefore, translations can be said to constitute a certain language variety with its own characteristics. The majority of investigations on the features of translations deal with the dominant languages, such as English (Baker, 1993, 1995, 1996; Laviosa, 1997, 1998), German (Kenny, 2001), or Chinese (Xiao et al., 2010). However, it remains unclear whether these features are common to the translations of other typologically different or minor languages. Research of translations should be especially relevant for less resourced languages as Lithuanian which are constantly affected by global languages and face the fact that in certain spheres, translations exceed original texts. For example, in Lithuania, in 2014, average editions of translated fiction were twice as large as the editions of original Lithuanian books (Markevičienė & Tamulytienė, 2015). Despite the prevalence of translations, there has been a lack of language resources representing translated Lithuanian. The aim of this poster is to present a compiled comparable corpus of original and translated Lithuanian (ORVELIT Lyginamasis originalų ir vertimų tekstynas). The 4-million-word corpus includes fiction and popular science texts and consists of four sub-corpora (original and translated fiction; original and translated popular science literature). Also, preliminary findings of corpus-based research of translated Lithuanian will be introduced with a specific focus on simplification and unique items hypothesis. The findings reveal that translated and original Lithuanian show differences in the type/token ratio, lexical density, and the distribution of pronouns and diminutives. It is believed that the created comparable corpus of original and translated Lithuanian will not only encourage further research of translations, but also have a practical value for trainee and professional translators, lecturers and language editors.

References


“Homo Austriacus” on the mountains. The discursive construction of mountaineering in an alpine heritage corpus
Claudia Posch (University of Innsbruck, Austria)

The Alps and mountains in general may not just be perceived as a geological entity but also as cultural constructs (Bubenhofer/Schröter 2012: 264). In accordance with Rak (2007: 111) we suggest viewing discourses about mountains and mountaineering as a rhetoric of their own, because they narrate very particular "conceptualizations, beliefs, evaluations and behaviors" (Bubenhofer/Schröter 2012, 265) members of a social group share. This social group of climbers and mountaineers as well as the activity mountaineering have become quite significant since the 18th century, especially in Central Europe. The discourses around this topic may have been “one of the key ways that modern, Western ideas about human activity, the idea of the body and the linking of gender to ideas about nationalism, colonialism, and race have been formulated” (Rak 2007, 112). This is because writing about mountaineering is equally important as the activity itself. Alpinist discourses offer climbers the possibility to prove their worth within their very elitist community, because they document the climb itself and provide climbing advice for anyone trying to repeat a route. Frequently they also have to provide some sort of justification for the activity itself – which without these discourses would be considered a very unnecessary risk. Expedition narratives, for example, are a way that mountaineers “talk” to or about each other, and so they provide much more than factual information. They are about social and ideological issues as well as about climbing, because they tell other climbers not only how to climb a certain mountain but how to be a climber too: very often they are about the ability and potential of the modern individual, about progress, conquest and aspiration. At the same time, those discourses entail utterances about people and cultures living in the mountains, who often are depicted as tradition-bound, even sometimes weak-willed. Mountaineering rhetoric hence produces a specific kind of climbing subjectivity which relates climbing to Western notions of identity. Therefore, mountaineering, a sport without spectators, relies heavily on discourses and for this reason was called “the most literary of all sports” (Barcott 1996, 65). Those discourses in the form of texts, speeches, movies etc. usually follow certain structures and patterns, because to be part of this community a certain stile has to be followed and specific topoi need to be activated. Lutz (2002) for example describes mountaineering as a condensed form of “masculine encounters with the body and nature”, which is not just “hypermasculine” but also soldierly and nationally coded. Topoi like ‘critique of civilization’, ‘individualism’, or the hero, who experiences and epiphany in the mountains are also quite generally known.

However, mountaineering and related discourses are very rarely treated in DA and there is little substantial research on the rhetoric and language of alpinism (especially in the German language), even though there is a significant number of texts on the topic. The proposed talk uses the corpus Alpenwort – Korpus der Zeitschrift des österreichischen und deutschen Alpenvereins (Corpus of the Austrian Alpine Club Journal) to investigate how the concepts and topoi mentioned above are linguistically constructed and linked with each other. Methodically we use a
'quantitatively informed qualitative' discourse approach, as suggested by Bubenhofer & Schröter (2012), Mautner (2012). Research questions are oriented along the lines of a thesis stated by Lutz: Heroism as something masculine that becomes visible in mountaineering as extreme sport, embodies virtues and values, which are often used to confirm identities of nations or even of a nationalism that reflects feelings of superiority towards other nations.

We investigate the following questions from a discourse-historical perspective:

- How are nations and people in the corpus discursively constructed?
- How is especially Austria constructed as the mountaineering nation par excellence?
- Which linguistic patterns and topoi are used to construct Austria as an "alpine super power" and the heroes as its soldiers?

We will especially focus on references/nominations (Wodak 2012), as well as argumentation patterns. The research uses a combined corpus-linguistic and discourse-historical approach to look for recurrent patterns of language use relating to these topics (e.g. n-grams, use of personal pronouns and nomination, predications, intensifiers).

To investigate these questions a large, linguistically annotated heritage corpus of alpinism will be used. This corpus contains 126 yearbooks of the Austrian Alpine Club Magazine (=Zeitschrift des Deutschen und Österreichischen Alpenvereins, ZAV) from the years 1869 to 1998. It consists of 42,000 book pages which result in approximately 18.6 million word forms and is available in a corrected and fully POS-tagged form. In its first decades the corpus reflects the ongoing touristic and cartographic exploration of the Alps. During the 20th century perspectives expanded to the mountains of the world. Topics such as environment and nature protection are discussed as well as questions of regional identity and cultural heritage. Thus, this corpus provides an excellent opportunity to investigate how concepts like heroism, femininity/masculinity, Western notions of identity, and ideas about nationalism, colonialism and race (e.g. "world mountains"), are discursively constructed.

References


Spoken BNC2014 EAS vs. the demographic part of the BNC:  
What can a study of tag questions tell us?  
Karin Axelsson (University of Gothenburg, Sweden)

The study of English spoken conversation has benefited tremendously from the demographic part of the British National Corpus (BNC) launched in 1994 (Burnard 2007). However, language change is, as always, going on so a corpus of spoken English reflecting the language of the early 2010s might give us a better understanding of how spoken English has changed over twenty years. This is the background to the development of Spoken BNC2014 (Love et al. 2017 fc.) to be launched in 2017. However, some researchers have already been allowed access to a subset. This Early Access Subset (EAS) has about 5 million tokens and is comparable in size to the demographic part of the original BNC. Henceforth, the EAS will be referred to as BNC2014E and the demographic part as BNC1994D.

BNC2014E appears to be suitable for the study of one of the most typical features of English conversation: canonical tags, as in It’s going to rain, isn’t it?, including non-standard innit derived from isn’t it/ain’t it (Andersen 2001:106). The term declarative tag question (DecTQ) is here used for the combination of a declarative anchor (e.g. It’s going to rain) and a tag (isn’t it, innit, etc.).

The aim of the present study of DecTQs in BNC2014E is to investigate their frequencies, formal features, functions and sociolinguistic features in contemporary English and make comparisons to their use in BNC1994D. After considerable random thinning of the search result in BNC2014E, 497 DecTQs were identified and analysed. Separate searches, without thinning of the results, were also made for the non-standard tags innit and ain’t it as well as for TQs with imperative anchors (ImpTQs). For BNC1994D, data from my doctoral dissertation (Axelsson 2011) was mainly used (1,315 instances) but some additional analysis of data from BNC1994D was performed for the present study, particularly as to sociolinguistic features. One hypothesis was that the tag innit is on the increase and that standard canonical DecTQs might therefore be on the decrease. There are indeed fewer DecTQs in BNC2014E than in BNC1994D (2,795 vs. 5,062 pmw) (without innit: 2,737 vs. 4,623 pmw), but the frequency of innit (including in it) has dropped even more: from 439 pmw in BNC1994D to just 72 pmw in BNC2014E (in the separate search for all instances of innit). These surprising results made me look closer at the comparability of the two corpora for the study of TQs.

Firstly, the two corpora have been compiled in different ways. BNC2014E uses crowd-sourcing where the requirements of good recording quality seem to favour more focused conversations than in BNC1994D, where the randomly selected respondents were told to record all spoken interactions during two days, i.e. also when the conversation itself was not the main activity going on. This difference is reflected in the fact that there are virtually no second-person ImpTQs (as in don’t tell her will you) in BNC2014E (1.4 pmw), i.e. much less frequent than in BNC1994D (40 pmw). ImpTQs deal with exchanging goods and services and not exchange of information (Axelsson 2011) and may thus tend to be more common if other activities than just talking are going on. The more focused conversations in BNC2014E may also result in fewer misunderstandings. The proportion of constant-polarity DecTQs (as in she’s working is she) is just five per cent in BNC2014E (vs.
almost ten per cent in BNC1994D). One reason to use constant-polarity DecTQs is to seek verification (Kimps 2007), e.g. if one questions what one believes to have heard another interlocutor say. Another consequence of the more focused conversations is probably a slightly higher level of style. An indication of this is that the proportion of reversed-polarity DecTQs with ellipsis of the subject and/or the finite in the anchor (as in *typical English weather isn’t it*) is clearly less common in BNC2014E (9%) than in BNC1994D (17.5%).

Secondly, BNC2014E is less well balanced for sociolinguistic features than BNC1994D, where the respondents were selected to reflect all ages, social groups and regions proportionally. As to the age of the speakers in BNC2014E, the category 19–29 is clearly overrepresented, whereas speakers younger than that are underrepresented. The fact that children below ten are almost absent in BNC2014E may also affect the language used by adults, as they would probably tend to talk differently to children than to adults. This may be an additional reason why second-person ImpTQs are rare in BNC2014E compared to BNC1994D: instances such as *Don’t touch anything will you* appear only in BNC1994D. The social grade groups with the highest normalised frequencies of DecTQs in BNC1994D are C2 and DE (significantly higher than for the rest of the corpus, i.e. grades AB, C1 and “unknown”). In BNC2014E, grades C2 and D are poorly represented (only about 2% each), whereas grades A, B and E contribute 24–31 per cent each. The facts that BNC2014E is not well balanced for sociolinguistic features, that the social grades of 38 per cent of the speakers in BNC1994D are unknown, and that social grade might affect the use of TQs complicate the comparison between the two corpora.

Thirdly, there are differences in the transcription principles. The low number of *innit* in BNC2014E might partly be due to the transcription guidelines saying “only use *innit* when you are sure: otherwise either use *isn’t it* or *ain’t it*” in combination with the clearer recordings required. Another problem is that the transcribers for BNC2014E were not allowed to use very much punctuation, practically only question marks, and that the size of the corpus is given in tokens (including quotation marks) instead of just words. For the calculations of frequencies and formal features above, the size of the two BNC corpora have been recalculated to exclude punctuation (BNC2014E then has 8.7 per cent more words than BNC1994D: 4,707,081 vs. 4,329,797). However, tokens are used in the calculations for sociolinguistics features as the sociolinguistic metadata is only supplied in tokens.

Fourthly, although there are as many as 376 different speakers represented in BNC2014E, some individual speakers have fairly large shares of the whole corpus. All the normalised frequencies presented above are actually calculated based on data from a subcorpus of BNC2014E, where all the utterances of 14 very prolific speakers (those with more than 75,000 tokens in the corpus) are excluded; in this subcorpus, there are 238 DecTQs uttered by 119 different speakers. The reason for this reduced dataset is that some speakers seemed to skew the results severely, in particular one speaker (an old man from Norfolk), who uses TQs to such an extent that, if the whole BNC2014E had been considered, his share would have been extremely high: 9.8 per cent of all DecTQs, 22.2 per cent of all instances of the tag *innit* and 45.1 per cent of all instances of the tag *ain’t it*. Among the 14 excluded speakers, the normalised frequencies of DecTQs range between 554 and 18,658 per million tokens. This very large variation in individual frequencies makes it vital for the study of TQs
that corpus material is spread over many individuals and that no individual is allowed to have a high share of the words/tokens.

It has been suggested to me that, instead of comparing frequencies of TQs per tokens/words, it would be more adequate to use frequencies per utterances, sentences/s-units, clauses or finite verbs. The number of utterances is similar in the two corpora (just one per cent more in BNC1994D). The number of s-units is given for BNC1994D but not for BNC2014E. The number of clauses is difficult to calculate in both corpora; however, it is possible to compare the frequencies of finite verbs, which is related to the number of finite clauses. The frequency of finite verbs (imperatives excluded) is about five per cent lower in BNC2014E, so calculations per finite verb would decrease the difference in the frequency of TQs somewhat between the two corpora. Interestingly, there are notable differences as to the distribution of finite verbs: finite forms of be and do are more common in BNC2014E than in BNC1994D, whereas it is the other way around for finite forms of have, modals and lexical verbs. In the case of have, this is reflected in a significantly lower proportion of DecTQs with forms of have in the tag in BNC2014E than in BNC1994D (the only statistically significant difference between the proportions of tag subjects and tag verbs between the two datasets).

The requirements of good recordings and the detailed transcription guidelines make BNC2014 relatively easy to use. The contexts of examples are much more coherent than in BNC1994D. The analysis whether a match is actually a TQ is facilitated by the fact that change of speaker is indicated already in the concordance lines with speaker codes, which also offer quick links to the sociolinguistic information. Despite the comparability problems described above, the endeavour to compile a corpus of spoken conversation from the early 2010s with comprehensive sociolinguistic information must be praised. Hopefully, the final version of Spoken BNC2014 will be somewhat better balanced for sociolinguistic features.

References


Words that go together: An exploration of the idiom principle in institutional spoken English
Adriano Ferraresi (University of Bologna, Italy), Silvia Bernardini (University of Bologna, Italy) and Maja Miličević (University of Belgrade, Serbia)

1. Introduction

The notion of the idiom principle was introduced by Sinclair (1991:110) to account for collocations, or “semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments”. Research in corpus linguistics and psycholinguistics has attempted to chart the extent of its applicability in different communicative settings and conditions, to understand the relationship between attestedness in corpora and formulaicity, or the property for a sequence of words to be “prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar” (Wray 2002:9).

2. Aims

Taking advantage of a unique communicative setting (the European Parliament), our aim in this paper is to investigate the extent to which different groups of English speakers, engaging in spoken monologic discourse under different task constraints, appear to resort to the idiom principle. Our analysis focuses on lexical collocations and attempts to answer two questions: 1) are there differences in the number of collocations used by different groups of speakers under different task constraints? 2) is there evidence of faster/slower access, or more/less holistic processing, in the different conditions, as indicated by prosodic features? It has been suggested that “[f]ormulaic sequences tend to be uttered with particular prosodic features such as alignment with pauses and intonation units, resistance to internal dysfluency [and] no internal hesitations” (Wood 2015:23), and that such features can be taken as cues to the holistic storage of formulaic sequences (Dahlmann and Adolphs 2007). Here we focus on disfluency signals (silent and filled pauses and false starts) within collocations as evidence in favour or against operation of the idiom principle.

Our corpus is a subset of EPTIC (Bernardini et al. 2016), including 80 speeches delivered in February 2011, selected so as to include an equal number of read-out speeches and speeches delivered impromptu by native English speakers and speakers of ELF, English as a Lingua Franca.1 Interpreted speeches (from Italian and French) are also included,2 leading to the setup shown in Table 1.

---

1 Even though MEPS have a right to speak in their native language, a few elect to speak English.
2 Including interpretations from two different languages was meant to limit the possible influence of the source language, while ensuring that a larger range of interpreters was sampled. No attempt to measure the source language variable was made in this particular study.
While the corpus is very small, it is highly comparable and apt to investigate authentic spoken English produced in a range of conditions. The comparison between impromptu and read-out speeches should highlight the different role played by the idiom principle in planned, written-to-be-spoken language vs. spontaneous, unscripted language (Erman and Warren 2000). The comparison between native and ELF speeches follows in the wake of extensive previous research on native vs. non-native (mainly learner) use of formulaic language (Kecskes 2007; Li and Schmitt 2010). Finally, simultaneous interpretations should allow us to observe if and how this condition of extreme cognitive effort, requiring a “roughly equal activation of two languages” (Sharwood Smith and Truscott 2014:208), impacts on the idiom principle.

3. Method

Collocation candidates are extracted based on corpus queries targeting the following structures:

- adjective+noun: e.g. fair elections
- noun+noun: e.g. eyewitness accounts
- verb+noun: e.g. using violence
- noun+verb: e.g. industry faces

We discard word pairs including proper nouns and numerals, and make sure that the remaining ones are syntactically well-formed (thus we retain the pair use feed in ”use of contaminated feed”, but discard touch authorities in “get in touch with the authorities”).

To evaluate the collocation status of the remaining pairs, frequencies are obtained from ukWaC (Baroni et al. 2009) and used to calculate word association strength relying on two association measures (AMs), t-score ($t$) and Mutual Information ($MI$): these are known to emphasize different types of collocations, i.e. highly frequent vs. strongly associated ones (Durrant and Schmitt 2009). The cut-off point between collocations and non-collocations is based on the median of the bigram scores: $MI \geq 3$, and/or $t \geq 8$; bigrams with frequency <3 are excluded (cf. Evert 2008).

Disfluency signals within the selected collocations are identified as occurrences of silent and filled pauses, which are transcribed in EPTIC speeches as “...” and “ehm” respectively, as well as false starts (e.g. “sustainable m- management”). Prior to collocation extraction, the reliability of pause annotation was independently checked against the audio files by at least two authors.

Two kinds of statistical analyses are performed. In the first, the number of bigram tokens scoring high on a single AM (high-$t$, high-$MI$) and on both AMs (high-$MI$&$t$) is calculated for each speech and expressed as a percentage (e.g. of high-$t$ combinations relative to the number of word combinations found in a speech). Percentages of each type of collocation are then used as an outcome variable in three sets of comparisons, which

<table>
<thead>
<tr>
<th>Subcorpus</th>
<th>Native</th>
<th>ELF</th>
<th>Interpreted from FR</th>
<th>Interpreted from IT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Impromptu</td>
<td>10 (2,352)</td>
<td>10 (2,117)</td>
<td>10 (3,089)</td>
<td>10 (2,736)</td>
</tr>
<tr>
<td>Read-out</td>
<td>10 (3,398)</td>
<td>10 (5,366)</td>
<td>10 (2,959)</td>
<td>10 (2,475)</td>
</tr>
<tr>
<td>TOTAL</td>
<td>20 (5,750)</td>
<td>20 (7,483)</td>
<td>20 (6,048)</td>
<td>20 (5,211)</td>
</tr>
</tbody>
</table>

Table 1. Subset of EPTIC used, with n. of speeches (and n. of words) per subcorpus.
are tested for significance using Wilcoxon rank sum tests: a) original vs. interpreted speeches, b) originals produced by native vs. ELF speakers, and c) originals delivered as read-out vs. impromptu speeches. In the second analysis, the presence vs. absence of a pause is used as a binary outcome variable in a logistic regression model with speech status (original/interpreted) and AM status (high-$t$/high-$MI$/high-$MI$&$t$) as categorical predictors. All analyses are carried out using R.³

4. Results

All comparisons returned non-significant differences, with one exception: read vs. impromptu speeches differ in terms of the percentage of high-$MI$ combinations ($W=117.5$, $p<.05$). When the analysis is performed separately for the native and ELF groups, the difference is only significant for the natives ($W=21.5$, $p<.05$; Figure 1).

![Figure 1. High-$MI$ collocations in native and ELF speeches.](image)

The regression analysis of disfluency signals (Figure 2) shows that, as might be expected, speeches interpreted into English contain more pauses overall than speeches originally produced in English. More interestingly, in both subsets of speeches the more robust collocations, those with high $MI$ and high $t$-score, are less likely to contain a pause or a false start than those scoring high on a single measure. These predictors (AM and speech status) contribute significantly to the presence of disfluencies (coefficients in Table 2).

³ [http://www.r-project.org/](http://www.r-project.org/)
**5. Discussion and conclusion**

This paper reports on a study focusing on the operation of the idiom principle in a small-scale yet closely comparable corpus of English speeches delivered at the EU Parliament by different sets of speakers and under different task conditions (read vs. impromptu, native vs. ELF, original vs. interpreted).

Our results concerning the number of collocations used suggest that ELF speakers in this international setting and specialized register do not differ significantly from native speakers in their use of both frequent and salient (high-$MI$) collocations, when improvising their speeches. These results would seem to lend partial support to Morgan’s (2014:63) hypothesis that “$MI$ sensitivity is not a marker of the so-called “native-speaker”, but rather a high degree of proficiency with a particular register”. However, the fact that native speakers use significantly more high-$MI$ collocations in their prepared speeches than in those delivered impromptu, while no such difference is found in the ELF speeches, is coherent with the repeatedly observed higher sensitivity of native speakers to strength of association (Ellis and Simpson-Vlach 2009). At least as concerns our setting and the lexical collocations we have focused upon, the native speaker preference for strongly associated collocations seems to be more related to their “functional utility”, than to the greater working memory demands associated with speech constructed in real time (ibid:62-3). This finding seems confirmed by the lack of significant differences between speakers and interpreters, despite the latter group’s much heavier cognitive load (also signalled by the higher number of disfluencies in interpreted output overall). Our results concerning disfluencies within collocations lend support to the idiom principle hypothesis, since the more robust collocations, i.e. those that are both frequent and strongly associated, are delivered with fewer hesitations by all groups.

In further work we would like to enlarge the corpus and repeat the analyses for collocation types rather than tokens, to take into account the possibility that native
speakers use a wider variety of collocations than ELF speakers (Granger 1998). Finally, access to the data analyzed here will be provided through the NoSketch Engine platform. Speech-to-video alignment is being performed, giving access to the synchronized videos from concordance lines, an especially welcome feature for corpus studies of spoken language.

References


Strategies for incorporating findings from corpus linguistics studies into the teaching of spoken English for business
Clarice S. C. Chan (The Hong Kong Polytechnic University, Hong Kong)

Corpus linguistics studies have shown the importance of interpersonal and pragmatic markers in spoken discourse (Carter & McCarthy, 2006; Handford, 2010; Koester, 2006, 2010). When interacting with people in the workplace, business English learners need to be aware of the relational and attitudinal functions performed by these markers (Chan, 2017). However, these markers and their functions are not adequately covered in published teaching materials for business English learners (Chan, 2009a, 2009b; Williams, 1988). To sensitize learners to features of workplace discourse and the role that some interpersonal and pragmatic markers play in mitigating face threats and showing solidarity in the workplace, findings from corpora of spoken business English, such as those from the Corpus of American and British Office Talk (ABOT) (Koester, 2006, 2010) and the Cambridge and Nottingham Business English Corpus (CANBEC) (Handford, 2010), can be applied to the teaching of spoken English for business. What business English teachers need are ideas for effectively incorporating findings from corpus studies into learning activities.

While various ideas that make use of written corpora in business English teaching have been suggested (see, for example, Chan & Frendo, 2014), there is a paucity of ideas for using spoken corpora or findings from corpus studies in the teaching of spoken business English. In light of this, I suggest some ways to integrate classroom learning activities with corpus-based or corpus-informed materials on spoken business English. In particular, I give examples of tasks and activities that raise learners’ awareness of the functions of interpersonal and pragmatic markers, such as language analysis tasks that put together findings from different studies and peer observation tasks that help learners compare their own language with findings from corpus linguistics studies. Data collected through questionnaire surveys and role-plays are presented to show the effects of these tasks on learners’ motivation and the language they produce. Overall, the findings suggest that effectively incorporating insights from corpus linguistics studies into classroom teaching entails not merely presenting learners with frequency lists, concordance lines and extracts from texts, but also understanding where learners need most help and finding ways to relate the insights from corpus linguistics studies to learners’ language needs. This would help learners to appreciate the value of corpus data and encourage them to learn from authentic language.

References


Geolocating German on Twitter
Hitches and Glitches of Building and Exploring a Twitter Corpus
Bettina Larl and Eva Zangerle (University of Innsbruck, Austria)

About 16% of EU residents speak German as a native language, which makes it the most widespread language within the European Union. While German is the official language in Austria, Germany and Switzerland, the language differs widely in the three countries: German is a pluricentric language with three standard varieties: German Standard German, Swiss Standard German and Austrian Standard German. The official borders between Germany, Austria and Switzerland also form the boundary between the three language standards. Additionally, to those national varieties, there are multiple varieties on the regional and dialectal spectrum.

Languages, and thus Linguistics, have always been influenced by technological developments and new forms of media. Each new development has brought new methods and approaches of how language can or should be studied and explored. Because of easy access and informal communication methods, increasing numbers of oral markers are being incorporated into written language. This is often showcased on social media platforms such as Twitter. Every tweet includes language output in the form of short messages that can contain different regional markers. Tweets can be geolocated, which means these language outputs can be assigned to the geographic location they were tweeted from.

This paper explores and describes the process of building a geotagged Twitter corpus of German tweets and the exploration of a preliminary sample. To research questions like “Is there a connection between the language output and the geographic location tweets were sent from?” and “Could, for example, lexical varieties be allocated to a specific region by geolocation information provided in tweets?” we are building a Twitter Corpus. The Corpus contains tweets collected via the Twitter streaming API, using a binding box around the rough approximation of the Deutscher Sprachraum and re-filtering the results for Tweets sent within Germany, Austria, Switzerland and South Tyrol/Italy. The data was gathered over a period of 24 months and more than 60,000,000 tweets were collected.

In this paper, we show and illustrate the way from data to corpus and how we address various challenges along the way.

References


Bouvier, Gwen (2015): What is a discourse approach to Twitter, Facebook, YouTube and other social media: connecting with other academic fields? In: Journal of Multicultural Discourses, vol. 10, no. 2. 149-162.


Few registers have perplexed linguists more than narrative fiction. At times, fictional prose behaves like a historical or biographical narrative, and at other times it creates fictional people who interact – often through fictional speech represented in a variety of forms (e.g. Semino & Short 2004). Fiction is typically seen as a linguistic variety shaped by creative use of language. As a consequence, findings from fiction do not always fit general patterns and might be discarded as outliers or idiosyncratic usage. Equally, when the focus is on patterns shared across the register, it seems fiction displays features that are less striking or contrasting than features of other registers. De Haan (1996, p. 38) observes: “fiction takes sort of a middle position between more formal writing on the one hand, and face-to-face conversation on the other”. One plausible explanation for this pattern is that a novel is made of different discourse levels (e.g. fictional speech and narration), but linguistic counts are taken across the whole text. When non-fictional registers are compared, Biber’s (1988) Dimension 1 (Involved versus Informational Production) very clearly shows a distinction between Face-to-Face Conversations, \( M = 35.3 \), and the narrative register of Biographies \( M = -12.4 \). Assuming that fictional speech mirrors actual spoken language, these findings suggest that treating fictional speech and narration as a single text will simply produce an ‘average’ of the dialogic and narrative features and will not necessarily represent the linguistic characteristics of either. Nevertheless, few scholars have attempted to distinguish between fictional speech and narrative when analyzing linguistic patterns in novels. This is to some extent due to a long-held belief that fictional speech is fundamentally different from ‘real’ spoken language (Page 1988) precisely because it is part of the narrative text as a whole.

Only a small number of studies have focused exclusively on fictional speech (e.g. Burrows, 1987; De Haan, 1996; Hubbard, 2002, Oostdijk, 1990). Others have analyzed samples of fictional speech for the purpose of estimating the characteristics of speech in time periods that pre-date audio recording devices (e.g. Biber & Finegan, 2001; Culpeper & Kytö, 2010; Kytö, Rudanko, & Smitterberg, 2000). Some recent research has focused on analyzing what Lambert (1981) calls the ‘suspended quotation’, or the interruption of fictional speech by narration e.g. Mahlberg, 2013; Mahlberg & Smith, 2012; Mahlberg, Smith, & Preston, 2013). Although there have been calls to reconsider the status of fiction as a single register (see, e.g., De Haan, 1996, 38-39; Egbert, 2012, 189), no research to date has approached fiction in this way.

In the most basic sense, the text of a novel can be divided into two parts based on formal features: fictional speech - within quotes - and narrative - outside of quotes (Mahlberg et al. forthcoming). In order to get a better understanding of the linguistic features of fictional narrative, the goal of this study is to carry out a comprehensive
linguistic description of the style of Charles Dickens, with a specific focus on the differences between narrative and fictional speech. Specifically, we aim to answer three questions in this paper:

1. In what ways does Dickens’s style differ between narrative and fictional speech?
2. In what ways does Dickens’s narrative and fictional speech change over time?
3. Can the narrative and fictional speech in Dickens’s novels be grouped in a meaningful way?

Methods

In this study we used a corpus comprising all 15 novels written by Charles Dickens (DNov), each of which was divided into two texts, one that contains all of the fictional speech (text inside of quotations) and one that contains all of the narrative (text outside of quotations), for a total of 30 texts. The DNov corpus contains a total of 3.8 million words of running text, with about 2.5 million words (66%) of narrative and 1.3 million words (33%) of fictional speech.

Each of the 30 texts in the DNov corpus was tagged using the Biber Tagger and processed to calculate normed rates of occurrence for 150+ linguistic features using Biber’s TagCount program. We used these counts to compute scores for each text on four dimensions from two previous Multi-Dimensional analyses. The first dimension comes from Biber (1988), and will be referred to as Biber_D1 (Involved versus Informational Production). This dimension was based on a corpus of texts from the LOB and London-Lund corpora. The other dimensions used in this study come from a description of stylistic variation in the FABLE (Fiction of America and Britain from the Late Eighteen hundreds) corpus (Egbert, 2012). The three dimensions from that study are FABLE_D1 (Thought Presentation versus Description), FABLE_D2 (Abstract Exposition versus Concrete Action), and FABLE_D3 (Dialogue versus Narrative). Dimension scores for each text were calculated by standardizing the linguistic counts for all relevant features using the z-score formula based on means and standard deviations from the corpora used in the Biber and FABLE studies.

In order to answer RQ1, we compared the mean dimension scores for the speech and narrative sub-corpora along each of the four dimensions. These comparisons were made using t-tests to test for statistical significance and Cohen’s d as a measure of effect size. RQ2 was answered using Pearson’s correlations between time (year of publication) and each of the three dimension scores from the FABLE study. The narrative and speech sub-corpora were analyzed separately. This yielded 6 correlation coefficients (3 dimensions x 2 text types). RQ3 was answered using a cluster analysis of the thirty texts based on scores for the three dimensions in this study. Hierarchical clustering was used to determine the ideal number of clusters to extract, after which k-means clustering was used to determine cluster membership of each text (see, e.g., Staples & Biber, 2015).
Results

The analysis for RQ1 resulted in significant differences and extremely large effect sizes between speech and narrative along Biber_D1 ($p < .0001$, $d = 9.47$), FABLE_D1 ($p < .0001$, $d = 6.49$) and FABLE_D3 ($p < .0001$, $d = 14.44$). FABLE_D2 on the other hand showed no significant difference and a very small effect size ($p = .59$, $d = .20$). These results reveal that Dickens’s speech uses more linguistic features associated with involvement (Biber_D1), thought presentation (FABLE_D1) and, unsurprisingly, dialogue (FABLE_D3). Dickens’s narrative, on the other hand, uses more features associated with an informational focus (Biber_D1), description (FABLE_D1), and narrative (FABLE_D3).

Figure 1 contains boxplots that display each of these differences.

In answer to RQ2, there were no statistically significant or noteworthy correlations between time and dimension scores in Dickens’s fictional speech. While there is a great deal of linguistic variation in Dickens’s fictional speech, this cannot be attributed to the variable of time using the linguistic measures used here. In the narrative sub-corpus, on the other hand, there was a strong and significant negative correlation between time and FABLE_D2 ($r = -.82$, $p = .0002$). This shows that more than 67% of the variance in FABLE_D2 scores can be accounted for by the variable of publication year (see Figure 2). This shows that over time Dickens’s narrative prose used more features related to concrete action and fewer features related to abstract exposition.
The cluster analysis revealed two clear top level clusters, with 15 texts in each that correspond perfectly to the narrative vs. speech categories. There are five clusters on the next level of the hierarchy that reveal additional groupings within the narrative and speech clusters. These reveal that although the most important predictor of Dickens’s style is the speech vs. narrative distinction, there are other stylistic features that can group his novels in meaningful ways.

**Conclusion**

This study represents an important step forward for research in corpus stylistics and literary studies. We show stark differences between the narrative prose produced by Dickens and his representation of character speech. We also reveal a diachronic change in Dickens’s narrative writing style that is not found in his fictional speech, suggesting that this pattern could not have been revealed had we not divided the novels into narrative and speech. Finally, we show that while the narrative-speech distinction is the strongest predictor of linguistic variation, there are other sub-clusters among Dickens’s novels within the clusters of narrative and speech that we plan to explore further.

The approach we have adopted here of analyzing narrative and fictional speech separately offers much to the study of novels, a register that has presented major challenges to corpus linguists and literary scholars. Our results strongly suggest that the narrative-speech distinction has to be accounted for in future research, especially quantitative corpus stylistics. This can be accomplished using our method of treating novels as two texts rather than one, or by some other means. This paper also reveals
new insights into the complex literary style of Charles Dickens and has wider implications for the conceptualization of narrative fiction.

References


Frequency and sequence: highest, first – lowest, last
Michael Barlow (University of Auckland, New Zealand)

The traditional approaches to the ordering of linguistic elements can be summed up as follows. The discourse perspective on sequencing is typically treated in terms such as given-new or topic-comment. Written or spoken output is about something and the sequencing has been seen in terms of a topic or theme, which typically occurs in initial position, and a following comment or rheme. In spoken discourse, there may be an expression in initial position that acknowledges the previous contribution. In addition, for each language there is a set of conventions or syntactic constraints that determine the kinds of structures permitted and hence ordering. Thus in English, the default order is Subject-Verb-Object and there are constraints of various kinds: adjectives precede nouns, and so on.

The influence of discourse and grammatical constraints is clearcut. In this paper we investigate ordering from another perspective. We can assume that there is a trend for high accessibility words to occur before low accessibility words but quantifying accessibility is a difficult task and so in this research we focus on frequency. Using a corpus of newspaper articles and a corpus of spoken American in a professional setting, we examine the frequency of words in different positions in sentences/utterances. That is, we investigate the frequency or rank of a word in first position in a sequence, compared with second position, and so on.

If we ignore context for the moment, we might expect a sequence to start with high frequency words and progress to lower frequency words as the sequence progresses. The higher frequency of the initial elements means that they are more accessible for the hearer (and also the speaker). We can think of this as a lexical accessibility.

There are different approaches to determining the frequency of each word in a sequence. It would be possible to use raw frequency but this measure will give an exaggerated view of the distance between words. An alternative is to use rank or log rank as the starting point since this reduces the variation between words of different frequency and so smooths out some of the variation.

The next question concerns which frequency list to use. An initial thought might be to use the BNC as a large corpus with general coverage. However, the generality of the corpus is not so essential and may be problematic when dealing with a single genre such as news stories from a single newspaper. Therefore the frequency list used is taken from the corpus being investigated and it seems reasonable to assess the frequency distribution of words within sentences by looking at their occurrence in the corpus as a whole. Thus, a frequency list is generated for the corpus and from that the rank for each word is used as an indicator of accessibility. Hapax legomena are not included in the frequency list.
The two corpora used are a Times newspaper corpus and a selection of files related to committee meetings from the Corpus of Spoken Professional American English. The latter consists of transcripts created by professional transcribers rather than linguists and so contains little in the way of prosodic features. The spoken usage is represented using sentences rather than prosodic units though with interruptions and false starts indicated by "—".

Since the number of words in a sequence varies, it is necessary to extract sentences/utterances of a set length so that we can get an overall picture of the frequency distribution in sequences of different lengths. For the spoken corpus, the sequences range from two words to ten words and for the written corpus, the ranges is 8- to 36-word sentences. The rank of each word in the sequence is provided. (The data is extracted and processed using python scripts.) Since turn-initial and non-turn-initial utterances may differ in their frequency profile, they are distinguished. An example of an 8-word turn-initial utterance is I don't know if it's possible or not. The rankings for each word in the sequence are: 6, 43, 39, 29, 28, 271, 33, 24. A non-turn initial sequence is: the whole licensing mechanism is still under discussion, which has the rankings: 1, 183, 524, 559, 8, 202, 294, 197.

Since the positioning of very frequent words such as articles may bias the results, the data is also processed with the most frequent words omitted.

If we examine the two sets of rankings above, we see that it is not the case that high frequency words always precede low frequency words. We do find in these instances, that the first word is the most frequent. It is to be expected that there will be a lot of variation in individual sequences but we wish to get a view of the general trends and do this we take the median rank value for each position in each sequence (8 words, 10 words etc.).

The results are remarkably consistent overall. For all the sequences examined in both the written and spoken corpora, we find that using a median value, the first word is always the most frequent and the last word is always the least frequent. This is the major finding. There are some more subtle variations in the frequency trajectory in different types of sequence that can be explored further.
The graph in Figure 1 provides an example of these findings. We see clearly the first word effect and last word effect, which holds even if the most frequent words are omitted. We also see that beyond these boundary points, there are patterns in the frequency trajectory. We do not get a straight line from high to low frequency. We can examine these patterns for both the written and spoken corpora and for sequences of different length.

As noted above, there are well-known constraints on the ordering linguistic elements. English has fairly strict grammatical word order constraints and there are discourse processes relating to information flow. Despite these constraints, the results described here show that there are also some frequency effects, which can be seen when viewing the median rank values for sequences of the same length.

**References**

In this paper I argue for a corpus-cognitive approach (cf., Mahlberg et al., 2016; Stockwell & Mahlberg, 2015) to address first the question of how discourses are related to each other, and second what they function is in context. To this end I first explore the potential of collocations and collocation networks to identify particular discourse relations and track their development. Second, I use frameworks from Cognitive Grammar (Langacker, 2008) to evaluate how discourse relations are expressed in context and how attention is distributed between related discourses to make a specific argumentative or ideological point.

Collocations are centrally connected to meaning-making, embodying implicit aspects of meaning such as connotations, or assumptions, and offering semantic analyses of a word in general or in a particular discourse. From a critical perspective, collocations can be informative of ideological attitudes, which is exemplified by the growing body of research that identifies these via collocational patterns. Yet discourses and their ideologies are not singular phenomena, but are complex, intertextually connected to each other, and continually negotiated (Sunderland, 2004). There is not just one but many discourses on a subject, whose relations can be visualised as a netted structure which results in ideological viewpoints. Corpus methods are particularly suitable to investigate discourses. Recent research on collocations and collocation networks (Baker, 2016; Baker & McEnery, 2015; Brezina, 2016) and the tools LancsBox (Brezina et al., 2015) and DiaCollo (Jurish, 2016) are promising ways into the research area of discourse relations. Such connections and the importance of the intertextuality of texts, has been explicitly identified in both discourse analysis and corpus linguistics as a major aspect of discourse that needs to be researched in more depth (e.g., Fairclough, 1992; Teubert, 2005).

Discourses as textual expressions also need to be held and negotiated cognitively and are the interface between the cognitive representations and processes on the one hand, and the particular societal position they express on the other hand (van Dijk, 1995). The relation between cognition and text is not direct, rather the repeated patterns in text provide familiar representations of people and events (Stubbs, 1996). These habitual ways of representation and argumentation do not determine the discourse, but impede a critical assessment of their underlying conceptualisation (cf. Hoey, 2005). I argue that certain research questions, such as how discourses are argumentatively related and understood by readers, can only be answered by considering both the textual and the cognitive aspect of discourse.

The discourse investigated in this paper is the debate on the legislation of abortion in the Republic of Ireland and Northern Ireland. The corpus is a sub-section the specialised Irish Abortion Debate Corpus (IADC), which comprises Irish online and offline newspaper articles from 2005 to 2016 on the issue of abortion. The subsection analysed in this paper is related to the death of Savita Halappanavar on 28 October 2012 and contains articles from October 2012 until October 2013; it comprises 6,711 articles with more than 3.5 million tokens. Halappanavar died due to the complications of a septic miscarriage after her request for an abortion was denied by hospital staff. Her death became a headline news story, leading to
national and international protests, increased media discussion on abortion access in Ireland, and was ultimately responsible in part for the Protection of Life During Pregnancy Act 2013, which allows abortion in the case that a woman’s life is at risk. Argumentative discourses, such as the debate on the legalisation of abortion, are particularly amenable to an analysis of discourse relations and how they form ideological messages because they necessarily harbour conflicting viewpoints. Many of the textual contributions to such discourse have the aim to persuade an addressee to accept a particular viewpoint, which results in drawing attention to some discourses over others.

In the analysis, I first identify the discourses connected to specific node words, such as mother, woman, or church, using collocation measures (MI3). All collocates are then sorted into ad-hoc created thematic groups. Only after the collocates have been identified and classified is the corpus separated into texts that contain two specific node words, which are assumed to influence each other ideologically, such as mother and church, and another set of texts where each node words occurs without the other. The thematic collocational behaviour of the two node words in these two sub-corpora is then compared to identify discourse relations between the nodes. To assess the functionality of these discourses, I further implement notions from Cognitive Grammar, mostly the trajector/landmark alignment and the reference-point relationship. This first explains the kinds of conceptualisations that are encoded by text-producers and how these are proffered for readers’ cognitive attention during interpretation. It second gives insight into how the discourses functionally relate to each other.

Such a combined corpus-cognitive analysis offers insight into how intertextual discourses are formed contextually and how they can reflect the reproduction of ideologies. For example, there is a clear association in the discussion of Halappanavar’s death whereby a mother is conceptualised as part of a mother-child entity and consistently attentionally de-focused in the text in comparison to the child. Extending the trajector/landmark alignment from sentences to discourses, I demonstrate that the ubiquity of this discourse in the Irish abortion debate at the time can be traced to one specific text and its influence, but also to how it was conceptually made available by this text and then subsequently repeated in other articles.

I argue that a consideration of how readers comprehend discourses and their relation to each other enriches our understanding of how discourse themselves: both how they are formed in context and how they can be emblematic of particular viewpoints. A viewpoint may be more persuasive because its is integrated into a coherent web of discourses which mutually support each other and whereby more attention is placed to some part of the web over others. In this paper I hope to demonstrate that corpus and textual analysis can profit from implementing frameworks from Cognitive Linguistics, while the empirical evidence of corpus methods can challenge and extend these frameworks.

References

Baker & T. McEnery (Eds.), *Corpora and discourse studies: integrating discourse and corpora* (pp. 244-265). London: Palgrave Macmillan.


The aim of the present paper is to report the findings related to the use of hedging and boosting resources by non-native novice writers of English in their argumentative essays (drafted and final versions) that was compiled as a special learner corpus during the academic year of 2015-2016. As Tono (2016) highlighted, learner corpora can play a significant role as “a powerful pool of documenting language learning progression” (p. 48) if we compile corpora efficiently. In the present study, after compiling learner corpora following Granger (1994), we aimed to explore hedging and boosting resources employed by Turkish learners of English, who use English as a foreign language, while writing an essay by following a process writing approach (Badger & White, 2000) with methodological and pedagogical contributions. We specifically focused on raising the awareness of the modality markers among L2 learners, i.e., hedges and boosters, as interactional resources (Thompson, 2001) using learner corpus given that the function of interactive resources through which the reader interpret the text by means of the interaction of the text and reader (see Hyland, 2004).

To create reliable and efficient sampling, we considered the following empirical corpus design issues: (i) We used the design criteria to build corpus that was based on learner and task variables (see Granger, 2003); (ii) we concentrated on the ‘representativeness’ through paying attention to the range of genres included (McEnery, Xiao & Tono, 2006) and (iii) we also updated the corpus for ‘permanence’ (Hunston, 2002) with the texts from the same group of learners, which can be considered as the methodological contribution of the present study. We want to highlight that the permanence will maintain for the sustainability of the corpus in the same research setting. The approximately 145,000-word learner corpus with two sub-corpora included (1) 90 very first composed versions and (2) 90 very last revised versions of argumentative essays written by the same pre-service English language teachers (age range from 19 to 25 years old) who study English in an EFL context in Turkey. 90 learners, whose proficiency levels in English were self-evaluated between intermediate and upper-intermediate, were assigned to produce written argumentative essays. Such a specialized corpus was purposefully intended for the sake of “giving insights into patterns of language use in particular settings” (Koester, 2010, p. 67).

The sub-corpora of the research were analysed by using WordSmith Tools 7.0 to see if there was a change in the usage of hedging and boosting resources by the learners between their very first versions and final versions of argumentative essays. The results of the analysis based in the first versions showed that Turkish writers of English employed more boosters in their drafted versions than hedges. However, it is interesting to note that the later analysis of the final versions of the same argumentative essays indicated a decline in the use of boosters after the explicit instruction of hedges and boosters in the classroom by providing examples from the drafted versions of the learners. This can be linked to a transfer issue from their L1 based on the claim of Akbas (2014), who suggested that Turkish L1 writers tend to employ...
more boosters in their writing. Considering the fact that novice writers of the study did not have enough experience with the traditions of writing in English, they might have applied their previous knowledge to the writing in the target language. This was also confirmed with the written feedback received from the learners at the end of the academic year. On the other hand, it was observed that the analysis of the final versions were more fruitful in terms of the hedging resources.

Although a number of studies suggested that corpora can be a valuable resource to produce classroom materials (e.g., Cobb, 1997; Cullen & Kuo, 2007), surprisingly, corpus-based activities still have not completely reached the classroom (see Gavioli, 2005; Römer, 2006 for criticisms) particularly for assessments in the mainstream classrooms. To contribute to the literature, we prepared materials from learner corpus of our research in Turkish context to raise awareness on hedging and boosting as a part of Advanced Reading and Writing course syllabus as course materials and assessment in the classroom. Also, we collected written feedback from learners anonymously to find out the efficiency of use of learner corpus in the classroom. We will provide sample classroom materials and written feedback to show how learner corpus data could be better exploited considering the needs of learners in the mainstream classrooms. Pedagogical perspectives are provided for material design and classroom methodology.

References


Subject-specific lexical patterns in Irish post-primary school textbooks and implications for ESP
Stergiani Kostopoulou (Applied Language Centre, University College Dublin, Ireland)

The comprehension and production of subject-specific language found in school textbooks can be a major challenge for students for whom the language of schooling is a second language.

An extensive body of research into the nature of the English language reveals that it is phraseological (Sinclair, 1996; 2005; Partington, 1998; Hunston and Francis, 2000; Stubbs, 2001), relying heavily on lexical patterns, or multi-word combinations that appear more frequently than expected in a particular linguistic register. These recurrent word strings appear to have multiple functions in subject-specific registers; they constitute fundamental building blocks of academic discourse (Biber et al., 1999; Biber, 2006) and contribute to the internal coherence of its structure (Hyland, 2008: 4). They further shape context-specific meanings (ibid.: 5) which means that each subject-specific register can be characterized by a distinct set of recurrent word sequences (Scott and Tribble, 2006; Biber and Barbieri, 2007: 265; Hyland, 2008). Thus, the description of high frequency word sequences can offer ‘insights into important aspects of the phraseology used by writers in specific contexts’ (Scott and Tribble, 2006: 132). In the light of these findings, it can be argued that helping students to understand and use subject-specific lexical patterns in context can facilitate access to subject-specific registers.

The present research was conducted in the context of the English Language Support Programme of the Trinity Immigration Initiative (Trinity College Dublin), which aimed to enhance the provision of English language support for immigrant students in Irish post-primary education.

This paper discusses the 100 most frequent 4-word clusters manifested in five specialized language corpora that were built based on 24 commonly used textbooks of English, History, Geography, Mathematics and Science used in Irish lowersecondary education. WordSmith Tools (Scott, 2004) were used to conduct quantitative and qualitative analyses. Word clusters are discussed in terms of density, structure, complementation patterns (i.e. adjacent words/phrases), discourse meanings and functions.

The findings demonstrate striking differences in the nature and manifestation of 4- word clusters across the five corpora. In terms of density, there is a higher concentration of clusters in the corpora of the two more technical subjects of Mathematics and Science. This may be attributed to the fact that conventionalized language is used to discuss numbers, symbols and graphic displays of data, which make up a large proportion of mathematical and scientific discourse. These findings are in line with conclusions drawn from other studies of disciplinary variation of word clusters (e.g. Biber, 2006). The structural analysis highlights the nominalized and structurally incomplete nature of clusters, supporting the findings of Biber et al. regarding word clusters in academic prose (1999: 998-1036). The qualitative insights into discourse meanings and functions reveal that lexical patterning directly reflects the distinct thematic concerns and communicative needs of individual subject areas.
This adds further evidence to the existing research that shows that each register employs a distinct set of lexical bundles (e.g. Biber and Barbieri, 2007: 265), which can be thus considered reliable indicators of linguistic variation (Hyland, 2008).

The pedagogical implications of the above findings are briefly discussed in relation to ESP learning, teaching, materials design and testing for immigrant students trying to access the language of schooling.

This empirical research represents the first of its kind in the Irish context. More broadly, it responds to the calls for more applied corpus linguistics studies which are motivated by an interest in second language education (e.g. Coxhead, 2010) as well as to the need for more localised ESP corpora to examine additional language varieties. The study also responds to recent calls for applied corpus linguistics research specifically into the language of secondary education to support students in their school-based studies (e.g. Flowerdew, 2009: 345; Coxhead, 2010: 466).

References


An annotated video corpus of interactions dealing with the collaborative construction of fiction in games
Peter Menke (Paderborn University, Germany)

In this ongoing project we are assembling a video corpus of broadcasts of various Let’s Play (LP) channels on platforms such as Twitch and YouTube. In particular, we focus on presentations of board games, card games and tabletop role-playing games, rather than video games (which are the prototypical game genre presented in LPs). Such interactions typically contain data with the following characteristics:

- There is face-to-face interaction between two or more participants.
- The participants are involved in imaginative, immersive games that work by collaboratively creating and inventing fiction (Fine, 2002; Herbrik, 2011; Schmidt, 2011). As a consequence, participants produce utterances that are interpreted as “in character” contributions. This means that they are uttered “inside the story”, and, thus, need to be interpreted in the context of the fictional reality created by the participants. For instance, in a round of Monopoly, utterances about mortgages or auctions influence the game reality, but they do not have any financial consequence for the players in the real world.
- Apart from this, the participants also produce “out of character” utterances that belong to the normal level of reality (and, often, serve metacommunicative functions). Two characteristic variations are (a) communication within the group of players, and (b) communication directed to the remote audience of the LP setting.

At the moment, we restrict the corpus to one language (German). On the basis of the corpus we investigate how people use language when they play. For this, we investigate questions such as the following:

- How do players design their contributions for the different communication partners (fellow players, bystanders, (a)synchronous remote audiences, etc.)?
- How do the “in character” contributions differ from “out of character” contributions?
- How do players comment the game from an outside perspective? How do they separate these utterances from those belonging to the game? Do they produce explicit signals for ‘switching realities’?
- How do the players resolve conflicts, unclear points, and misunderstandings related to the rules of the game at hand?

We already worked on several of these questions in smaller case studies based on single LP sessions. As a next step, we assemble this corpus that will enable us to rely on a larger data basis. This growing corpus consists of videos that have been made publicly available on the respective video platforms. After obtaining permission for the download, we create outlines describing the phases of the game dialogues along with the main events, both on the game level and outside of it. The audio material is transcribed following the GAT 2 conventions (Selting et al., 2009) using EXMARaLDA (Schmidt, 2002, 2009) and annotated following an action-oriented annotation schema that is being developed and refined during this project. Our goal is to support qualitative approaches such as Conversation Analysis (as already performed in the case studies) as well as quantitative investigations at a later stage, where frequencies and durations of dialogue or game phases are analysed.
References

This presentation focuses on a corpus-assisted discourse analysis of interaction among Christians, atheists, and Muslims on YouTube. The internet and social media sites in particular, offer access to diverse audiences, but for religious users, conflict can occur when attempting to make videos which are viewed by communities of believers who share the same faith, at the same time as users who are openly hostile to their beliefs. This presentation aims to describe how users manage these competing pressures by looking closely at how they speak about their faith, and how they speak to others on social media sites, particularly through the medium of online video.

Taking a discourse-dynamics approach (Cameron, 2015) to video interaction, I show how interaction across different levels of discourse activity can be used to describe trends in presentation of belief among people of different faiths and no faith. I present a case study of responses to one particular Evangelical Christian preacher, Joshua Feuerstein, by a Muslim YouTuber and an atheist YouTuber, using a corpus of 67 video pages (including 6 hours and 47 minutes of talk and 60,888 comments). The project investigates how user talk differs depending on the context of interaction and how users accommodate diverse audiences. The analysis included two stages: positioning analysis (Bamberg, 1997; Harré & van Langenhove, 2008) and keyword analysis (Baker, 2006).

I first show how local, situated interactions create positionings that fit within larger storylines that serve each user’s own ideological position. The can be observed in the conflict between users, as users position one another and respond to positionings. In the central conflict of the data, the Christian positions atheists as aggressive, antagonistic, and ignorant for ‘believing in evolution’. These positions are constituent parts of a larger narrative that draws an equivalency between ‘faith’ in empirical evidence and faith in God. In the response, the atheist positions the Christian as ignorant, building a narrative where atheists are logical and sensible, and Christians follow archaic beliefs and are illogical. These competing narratives show how online ‘drama’ or antagonistic debate (Burgess & Green, 2008) provides affordances for users to create novel content and put forward their own ideological positions in response to those opposed to them. Every negative comment from another user is an opportunity to reassert one’s own narrative.

Building on positioning analysis, I then show how keywords in the whole dataset can be used to contextualise the interaction among users and show how positionings within individual videos are part of larger narratives that stretch across many videos. Keyword analysis is used to investigate how positionings uncovered in the micro analysis are true of the larger data, particularly in the responses to videos in text comments. Looking at the top 100 keywords in the sub-corpora of users’ spoken discourse and the text comments on each user’s videos compared to BNC reference corpora using WMatrix, I show how three distinct ways of talking about religion emerge, one in which atheist discourse focused on science and logic, while
the discourse surrounding the Muslim and Christian videos was oriented towards religious practice. The keywords also reveal the importance of Christianity in the discourse of all three users and those commenting on the videos, as well as different registers that emerge when users speak to different audiences. I show how the keywords can be used to return to the process of discourse analysis and further uncover how discourse among users develops over time.

My analysis shows how user interaction, particularly hostile interaction around issues of faith and belief, is affected in online contexts. I focus on three main findings. First, given the open nature of public social media interaction, users are compelled to respond to a broader social contexts and this engagement requires some adaptation at least in the presentation of belief. This affordance allowed users to produce content that drew praise from likeminded users and derision from opponents, leading to a rising profile on the site when they posted their videos. Second, the content and themes of the arguments are not especially unique and are the result of ongoing interaction among people of different faiths. Historical arguments for and against particular faiths and between different faiths are enacted in online spaces, sometimes with very little development in terms of their content or approach. Third, talk about religious issues which is driven by and oriented towards popular personalities did not seem to support the growth of tangible affiliation among users or communities. While this has been observed to be the case in other studies, the analysis shows that the particular users created a hierarchical relationship between themselves and the people that watched their videos. The result was a lack of evidence for a sense of shared experience, but rather ad-hoc affiliations between users who held common beliefs.

References

This presentation focuses on a corpus-assisted discourse analysis of interaction among Christians, atheists, and Muslims on YouTube. The internet and social media sites in particular, offer access to diverse audiences, but for religious users, conflict can occur when attempting to make videos which are viewed by communities of believers who share the same faith, at the same time as users who are openly hostile to their beliefs. This presentation aims to describe how users manage these competing pressures by looking closely at how they speak about their faith, and how they speak to others on social media sites, particularly through the medium of online video.

Taking a discourse-dynamics approach (Cameron, 2015) to video interaction, I show how interaction across different levels of discourse activity can be used to describe trends in presentation of belief among people of different faiths and no faith. I present a case study of responses to one particular Evangelical Christian Facebook preacher, Joshua Feuerstein, by a Muslim YouTuber and an atheist YouTuber, using a corpus of 67 video pages (including 6 hours and 47 minutes of talk and 60,888 comments). The project investigates how user talk differs depending on the context of interaction and how users accommodate diverse audiences. The analysis included two stages: positioning analysis (Bamberg, 1997; Harré & van Langenhove, 2008) and keyword analysis (Baker, 2006).

I first show how local, situated interactions create positionings that fit within larger storylines that serve each user’s own ideological position. The can be observed in the conflict between users, as users position one another and respond to positionings. In the central conflict of the data, the Christian positions atheists as aggressive, antagonistic, and ignorant for ‘believing in evolution’. These positions are constituent parts of a larger narrative that draws an equivalency between ‘faith’ in empirical evidence and faith in God. In the response, the atheist positions the Christian as ignorant, building a narrative where atheists are logical and sensible, and Christians follow archaic beliefs and are illogical. These competing narratives show how online ‘drama’ or antagonistic debate (Burgess & Green, 2008) provides affordances for users to create novel content and put forward their own ideological positions in response to those opposed to them. Every negative comment from another user is an opportunity to reassert one’s own narrative.

Building on positioning analysis, I then show how keywords in the whole dataset can be used to contextualise the interaction among users and show how positionings within individual videos are part of larger narratives that stretch across many videos. Keyword analysis is used to investigate how positionings uncovered in the micro analysis are true of the larger data, particularly in the responses to videos in text comments. Looking at the top 100 keywords in the sub-corpora of users’ spoken discourse and the text comments on each user’s videos compared to BNC reference corpora using WMatrix, I show how three distinct ways of talking about religion emerge, one in which atheist discourse focused on science and logic, while
the discourse surrounding the Muslim and Christian videos was oriented towards religious practice. The keywords also reveal the importance of Christianity in the discourse of all three users and those commenting on the videos, as well as different registers that emerge when users speak to different audiences. I show how the keywords can be used to return to the process of discourse analysis and further uncover how discourse among users develops over time.

My analysis shows how user interaction, particularly hostile interaction around issues of faith and belief, is affected in online contexts. I focus on three main findings. First, given the open nature of public social media interaction, users are compelled to respond to a broader social context and this engagement requires some adaptation at least in the presentation of belief. This affordance allowed users to produce content that drew praise from likeminded users and derision from opponents, leading to a rising profile on the site when they posted their videos. Second, the content and themes of the arguments are not especially unique and are the result of ongoing interaction among people of different faiths. Historical arguments for and against particular faiths and between different faiths are enacted in online spaces, sometimes with very little development in terms of their content or approach. Third, talk about religious issues which is driven by and oriented towards popular personalities did not seem to support the growth of tangible affiliation among users or communities. While this has been observed to be the case in other studies, the analysis shows that the particular users created a hierarchical relationship between themselves and the people that watched their videos. The result was a lack of evidence for a sense of shared experience, but rather ad-hoc affiliations between users who held common beliefs.

References

KonText – a modern, customizable corpus query interface
Tomáš Machálek (Institute of the Czech National Corpus,
Charles University – Czech Republic)

We present a fully featured corpus query interface based on open-source system NoSketch Engine. The aim is to present a customizable tool providing features that emerged from long-term feedback given by researchers and users of the Czech National Corpus (CNC).

To be able to introduce a fully operational service fitting the needs of our present-day users in a reasonable time, we based our approach on our extensive experience with the NoSketch Engine (NoSkE) corpus query engine which is maintained and developed by Lexical Computing Ltd. as a simplified, open-source version of their commercial software Sketch Engine (SkE). Although open-source, NoSkE properties and project direction are tightly coupled with its commercial counterpart making the possibility of its direct adaptation to our needs through community-driven development rather unfeasible.

Our solution is based on reusing the core part of NoSkE – the Manatee-open library and related utilities which provide essential indexing, searching and analytic functions. Such a decision ensures that KonText can operate on the same indexed data and provide the same analytic results as NoSkE while avoiding development of highly specialized and complex text search software. Our primary focus is set on the development of a new interface and additional services allowing advanced functions for both querying corpora and result presentation/manipulation. An integral part of the project is to design the application as a set of interchangeable building blocks communicating with each other via defined interfaces. This allows other institutions and individual developers to adapt KonText to their specific needs without rewriting its core. KonText is a fully operational and mature software deployed at the CNC since 2014 that currently handles more than 1,650 user queries per day (any further actions that operate on query results, e.g. sorting or filtering, are not included in this count).

KonText is a three-layer system. The first (bottom) layer is composed by different backend server services. Unmodified Manatee-open library plays the primary role here but there are also our custom functions (e.g. database-stored meta-data, calculation control) present there. KonText provides an increased emphasis on scalability which means that in case of growing user base, another server can be added for a better system performance.

The third (highest) layer is the interface itself which is loaded and executed in user’s web browser. The source code in this case has been mostly written from the ground up and remaining pieces of the original NoSkE code are gradually being replaced.

The second layer sits between the two and provides a communication between user interface components and backend services. The source code of this layer is derived from NoSkE with some modules intentionally kept with minimum changes (esp. the ones tightly related to Manatee-open library) while others heavily rewritten or added (request processing, custom data access).

From the end-user perspective, we try to keep the core user experience like the one provided by NoSkE. This makes the possible user transition easier as the users are already familiar with basic application operation.

Our additions can be divided into three categories. The first category contains functions related to query construction and data selection:
• a module for interactive text selection (based on a combination of criteria) that facilitates user creation of tailor-made subcorpora with the possibility to define also alignment-based subcorpora,

• a function allowing custom text type proportions in subcorpora allowing the users to define a balanced corpus according to their specific requirements,

• a fully re-editable query where any operation (query, filter, sorting, sample) can be changed while keeping the other applied operations intact,

• a local CQL (Corpus Query Language) parser – currently used as a syntax checker but also allowing further future improvements in interactive CQL editing,

• a convenient visual data-driven widget allowing users to select a required tag value (for positional PoS tag formats),

• a storage for recent queries allowing users to review their previous queries for later reuse with data stored on server making it available from any computer and accompanied by additional useful information (date and time, queried corpus, used query type and parameters),

• two alternative modules for finding and selecting a corpus to work with (search by keywords and description or by expanding a category tree).

The second category of KonText's additions and changes is related to result presentation and manipulation. These include:

• manually selecting concordance lines and attaching custom numerical labels to them. Such a selection can be further filtered, exported or passed to other users via a URL address,

• support for direct export to Microsoft Excel format,

• an extended support for spoken corpora:
  o regions of texts can to be accompanied by audio chunks for direct playback,
  o intuitive visualization of dialog structure in spoken conversational corpora,

• rendering of dependency syntax trees.

The third category contains user-transparent changes motivated by improving performance and modularity of the system. While almost invisible for a user, these changes often belong to the most arduous ones as they intervene the core functionality of the original NoSketch Engine code. These changes include:

• a rewritten asynchronous processing of concordance search which now allows distributing the function to multiple servers,

• an improved caching of frequency distribution and collocation result pages allowing a faster navigation between pages.

KonText is an open-source software licensed under GNU GPL 2 (same as NoSkE) and publicly available on GitHub at https://github.com/czcorpus/kontext. A running production version can be found at https://kontext.korpus.cz/. Its development is maintained by the Institute of the Czech National Corpus but it is open to other contributors as well. Recently, a cooperation on further development of KonText has been established with the Institute of Formal and Applied Linguistics (Faculty of Mathematics and Physics at Charles University) and KonText has been adopted as the query engine at the LINDAT/CLARIN repository. There is also a positive feedback from individuals that examine possibilities of deploying KonText in their (mostly academic) environment. KonText is in daily use by researchers and students across the Czech Republic and it is under an active development with continuous bug fixing and major updates approximately every six months.
References

Adapting the BNC for sociolinguistic research – a case study on negative concord
Susan Reichelt (ESRC Centre for Corpus Approaches to Social Science, Lancaster University, UK)

Introduction

The following paper details the use of a specialized sub-set of the BNC (BNC64) for sociolinguistic purposes. The corpus, consisting of close to 1.5 million words, is based on the demographic part of the main BNC and offers a balanced sample of speakers across social variables such as gender, age, social class, and regional background. The study is the first stage of a comprehensive re-purposing of the BNC.

Negative concord, the feature used as a case study here, is a non-standard and stigmatized variant that appears in contexts where two negations are used within a single phrase. Prescriptive language norms dictate this form as ungrammatical or illogical (where two negatives form a positive) and give preference to forms with a single negative marker. The paper focuses on categories of gender, age, dialect, and social class and adds to reports of negative concord using the BNC (Anderwald (2002), Muntañà (2008)), highlighting in particular sociolinguistic aspects and its methods used.

The paper is thus aimed at providing some insights into social distributions of negative concord using corpus data, as well as providing detailed considerations in preparing and adapting existing larger-scale corpora for sociolinguistic research.

Data and feature background

In order to obtain representative results from this analysis, it was decided to adapt speaker variables given in the BNC64 to allow for better representation across speaker groups. Thus, dialect areas were broadened from an initial 21 areas to five. Further, ambiguous speaker variables (as found in regional backgrounds and social class), were omitted altogether.

The updated corpus encompasses comparable spoken conversations of 48 speakers. The analysis represents social variation of negative concord across the modified corpus. Social categories that are provided in broad and narrow forms (class and dialect) were initially tested for broad distinctions. Following that, variation that was shown to be statistically significant was further tested in narrow distinctions.

Negative concord (or double negation, NC henceforth) appears in various syntactic contexts and, even though relatively easy to define from a given text, needs to be operationalized thoroughly for larger corpus data. The principle of accountability states that variants within a closed set should include not only the tokens that are of interest for the current study (NC), but also instances where the token could have appeared but did not (“zero” tokens). Zero tokens for NC are structures that are considered as Standard English forms – forms that are considered grammatical and do not carry stigma. In order to give an account that includes both
negative concord as well as the Standard English variant, the token extraction was limited to two rather specific types of NC. Both types are defined through a negated auxiliary (such as can't, won't, don't, etc.), a following verb, and either an indefinite noun (such as anything, anyone, nothing, none, etc.) or a negative particle (such as any, owt, no, nowt) in combination with a noun. Examples below show tokens as found in the BNC64 (provided with line number) and the alternative variant (given with *) in the two contexts included (structural delimitation).

(1) They can't suggest anything. (270)
   *They can't suggest nothing.
(2) Couldn't hear nothing for splashing. (440)
   *Couldn't hear anything for splashing.
(3) Because Sainsbury didn't have any yesterday. (51)
   *Because Sainsbury didn't have none yesterday.
(4) I don't want no cake. (19)
   *I don't want any cake.

Token extraction

Limiting the search terms for NC (and zero forms) to this specific syntactic pattern which can be described as neg.AUX + VERB + ind.NOUN allowed for relatively straightforward token extraction through the corpus analysis toolkit AntConc.

All tokens were coded for actual NC (and zero form) functions, as well as for syntactic elements. Every token was extracted including speaker information so that meta-data that was provided was easily lined up with the analysis data.

In addition to social factors (age, gender, social class), grammatical factors such as the kind of particles that combined into the construct, were also included in the analysis, though will not be the focus of the current paper.

The following table summarizes all tokens that were included in the main analysis. Approximately 20% of all negations following the pattern neg.AUX + VERB + ind.NOUN appeared as NC. This is higher than comparable studies showed (14.3% in Anderwald 2002:105).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Total</th>
<th>Standard</th>
<th>Negative concord</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>373</td>
<td>N</td>
<td>%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>296</td>
<td>79.36</td>
</tr>
<tr>
<td></td>
<td></td>
<td>77</td>
<td>20.64</td>
</tr>
<tr>
<td><strong>Age grouping</strong></td>
<td></td>
<td>296</td>
<td>79.36</td>
</tr>
<tr>
<td>Under 35</td>
<td>128</td>
<td>105</td>
<td>82.03</td>
</tr>
<tr>
<td>35 and above</td>
<td>245</td>
<td>191</td>
<td>77.96</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td>277</td>
<td>87.18</td>
</tr>
<tr>
<td>Female</td>
<td>261</td>
<td>199</td>
<td>76.25</td>
</tr>
<tr>
<td>Male</td>
<td>112</td>
<td>97</td>
<td>86.61</td>
</tr>
<tr>
<td><strong>Dialect</strong></td>
<td></td>
<td>259</td>
<td>85.73</td>
</tr>
<tr>
<td>North</td>
<td>56</td>
<td>42</td>
<td>75.00</td>
</tr>
<tr>
<td>Midlands</td>
<td>118</td>
<td>98</td>
<td>83.05</td>
</tr>
<tr>
<td>East</td>
<td>82</td>
<td>63</td>
<td>76.83</td>
</tr>
<tr>
<td>London</td>
<td>50</td>
<td>42</td>
<td>84.00</td>
</tr>
</tbody>
</table>
Results

For the multivariate analysis the Rbrul package that loads into R was used. Multivariate analysis results show that social class is the most telling social category in NC variation, followed by gender. Neither age nor dialect area seem indicative (in terms of statistically significant variation) of this non-standard feature.

Stat. values from the analysis:
Class (0.000686) + Gender (0.0151) + Age (0.196) + Dialect (0.735)

With social class being an important factor here, a following analysis included narrower descriptions for both classes. The lower class group is made up of C2 and DE, while the middle class consists of AB and C1.¹

1 If the use of negative concord is socially stratified with a higher relative use of the standard variant by the higher social class (and vice versa), the preference of negative concord would increase from highest (AB) to lowest (DE) class. The analysis however shows that the class DE (semi-,unskilled) is higher in their relative use of the standard variant than the next higher social class C2 (skilled manual).

Discussion

Previous studies, looking at individual speaker variables, have found there to be stratification in region (Anderwald 2002). It is not clear whether different results are due to different samples from the BNC or due to the inclusion of other factors in the present study.

A possible interpretation lies with an awareness of the feature by the lowest social class. As was seen with other non-standard features, lower social class members, or those that are on the cusp between two classes, would oftentimes adjust their linguistic behaviour upwards. While the data provided here only gives a limited amount of information about the speakers and their social backgrounds, this could serve as a possible explanation for the social class distribution.

In terms of how the unusual gender distribution maps onto social class variation shows that the middle class men do not actually use NC and only at relative low rates in lower class contexts. However, the expected decline of non-standard features with increasing social class can be traced. It is the lowest social class for the female speakers that is surprisingly low (which might be a reflection of the class analysis above where it might be the lowest social class female speakers that are consciously avoiding using non-standard stigmatized features). Taken together female speakers have a much higher rate of NC overall, and male speakers only overtake NC frequencies in the lowest social class.

¹ AB – Managerial, administrative, professional; C1 – Junior management, supervisory; professional; C2 – Skilled manual; DE – Semi- or unskilled
Further data is needed to investigate this distribution more thoroughly. In terms of age and dialect groupings the present data did not suggest that the use of NC over standard variants is correlating with differences between younger or older speakers, or where they are located within England (contrary to previous research).

Conclusion

Using larger-scale corpora for sociolinguistic studies offer a wealth of (new) information about language feature use, as well as stratification across speaker backgrounds. Unfortunately, efforts into preparing corpora for sociolinguistic analyses are quite high and many features do not lend themselves to corpus-based studies following variationist methodologies. The case study shows that even with well researched features such as negative concord, there is still room for more diverse approaches and methodologies. This is a first step of adapting existing data in order to create new research, enhancing the field of corpus-based variationist sociolinguistics.

Acknowledgement

The research has been supported by the ESRC grant no. EP/P001559/1.

References

A mountain of work. Building an Alpine Heritage Text Corpus
Claudia Posch, Gerhard Rampl and Bettina Larl (University of Innsbruck, Austria)

The proposed poster demonstrates the process of building a large heritage corpus in the German language, which is an ongoing project at the University of Innsbruck. Inspired by the Swiss project Text+Berg digital (Bubenhofer et al., 2015) the Austrian project Alpenwort’s is currently working on a POS & NER annotated corpus of alpine heritage texts. The Alpenwort corpus contains 126 yearbooks of the Austrian Alpine Club Magazine (ZAV =Zeitschrift des Deutschen und Österreichischen Alpenvereins) starting from as early as 1869 until 1998. All of the volumes were already digitized and OCRed and are now part of a TEI-conform XML corpus with approximately 18.6 million words. One particular problem the project had to face during the digitization process was the large amount of text in Gothic script (The ZAV was issued in Gothic script from 1914 to 1962). To solve the resulting OCR-errors a range of semiautomated correction steps were developed.

The ZAV is an extraordinarily interesting source because of its continuity but also because of its thematic diversity. In its first decades the magazine contributions reflect the ongoing touristic and cartographic exploration of the Alps and the economic and scientific discoveries involved. During the 20th century perspectives expanded to the mountains of the world. Globally relevant topics such as environment and nature protection are discussed as well as questions of regional identity and cultural heritage.

The proposed poster presentation gives an overlook of the important steps in our efforts in building the Alpenwort Corpus: From scanning more than 42.000 book pages to Optical Character Recognition to logical structure extraction, the correction of structural elements and OCR-correction. We will also discuss the automated data annotation and enrichment, that contains tokenization, POS-tagging as well as named entity recognition.

The Alpenwort corpus will be freely available for the research community later in 2017 as an XMLversion as well as integrated in the tool Hyperbase, which was developed by our project partners in Nice, France.

References
Frequency distribution and usage of discourse markers in English to Russian translation: learner vs. professional translational behavior

Maria Kunilovskaya (University of Tyumen, Russia)

The major motivation behind this research is understanding linguistic behavior of translation students in their mother tongue during translation. We hypothesize that translators’ linguistic choices are different from those made in the target language by native speakers outside situations of translation. At the same time these choices indicate whether translators can effectively cope with cross-linguistic problems and counteract known translational tendencies against the backdrop of the existing translational norm (Mauranen, 2004) represented by professional translations. Ultimately, the intensity of ‘being a translation’ can be used to describe textual quality of translation (Scarpa, 2006) and subsequently, characterize translator’s professional competence in text production. In this research this approach to the study of translational language is realized with regard to discourse markers (DM) distribution.

Levels of explicit text connectedness have been on the linguistic research agenda in computational and corpus linguistics for many years. It is an important textual feature that reflects peculiarities of text production. From the translational perspective, DM frequencies are used to establish differences between translations and non-translations and are interpreted as a linguistic indicator of several tendencies in translation such as explicitation, simplification and convergence (Olohan, 2001; Chen, 2006; Denturk 2012). Previous work shows that translation-specific DM frequencies are a composite effect of at least three major factors: 1) specific cognitive processes that lead to source language independent translationese hypothesized by Baker (Baker 1993) (simplification, explicitation); 2) interference from the source language which can be used to establish direction of translation and detect differences between parallel corpora (Cartoni et al., 2011) and 3) adaptation to target language norms, especially for language pairs with contrasting cross-linguistic frequencies of DM (normalization). This genetic complexity of specific DM frequencies in translations calls for complex corpus resources (including comparative and parallel components) to disentangle them.

We set out to reveal tendencies in translational behavior at different competence levels by describing the frequency distributions of two functional types of DM (connectives and epistemic commentary markers) in learner and professional translations against sources and non-translations. Connectives are limited to mostly parenthetical items such as that is to say, consequently, however, by the way (to give examples of the four major semantic groups and different morphological structures; the search list totals 119 items for English and 95 for Russian). The group of epistemic commentary markers includes adverbials and synonymous phrases of various morphological status that convey the meanings of doubt and certainty described in Biber et al (1999). The respective search lists count 103

---

1 This work has been partly supported by the Russian Foundation for Basic Research within Project No. 17-06-00107
We compare data from a parallel translational learner corpus (http://www.rus-ltc.org/) and a corpus of professional translations to customized genre-controlled selections from English and Russian national corpora. All subcorpora are limited to mass media texts. The total size of the research corpus amounts to 10 mln tokens. Using independent predefined lists of targeted items for each language, we explore cross-linguistic differences and their influence over the two types of translation.

This research tries to marry extent and depth to follow Hunston’s metalanguage for corpus research description (Hunston, 2012). It refines and extends the reach of our previous study (Kunilovskaya, 2017) by bringing into focus epistemic stance markers to supplement information on connectives. We use quantitative methods and comparative corpora to isolate translationally distinctive DM (Chen, 2006) – items that have statistically different frequencies in translations as in originals – and offer detailed analysis of the focused items and their groups based on parallel data to explain differences in frequencies from the translational point of view.

To demonstrate our quantitative findings we visualized frequency statistics of different semantic groups of connectives in Figure 1. It shows the ratio of the four semantic types of connectives in the parallel subcorpora of learner and professional translations against the reference corpora in English and Russian. It can be seen that (given our search lists) learner translators use more connectives, which reflects cross-linguistic differences and is indicative of normalization and explicitation.

Figure 1. Contrastive and comparative frequencies of connectives by semantic groups (from bottom up: elaboration, inference, contrast, sequence)
Looking inside bi-texts helps to reveal typical and legitimate translational shifts in types of connectives used as well as typical errors, associated with certain English DM. To exemplify results of qualitative analysis and patterns observed, we consider the case of na samom dele (на самом деле) below. It is one of the connectives significantly overused in learner translations (p < 0.0271 from a two-tailed Wilcoxon rank sum test run on vectors of normalized frequencies obtained from every text in both translational corpora and from the reference corpus). In our multiple parallel corpus containing 11K of English sentences and around 30K of their translations there are 117 occurrences of this connective. In half of all multiple translation units (38 out of 75) its use is triggered by the polyfunctional in fact. This English prompt is referred to the group of most frequent adversative adverbials by Liu (2008) (162.63 ipm) and is reported to predominantly signal the discourse relation of expansion (usually by way of corroborating the previous statement) in the PDTB (Pitler et al., 2008). Na samom dele has the relative frequency of 70.3 ipm in the Russian National Corpus. Its semantics is centered on the idea of actuality, with sentence-initial uses expressing contrast between the state of affairs alleged in the previous discourse and the perceived reality (Baranov et al., 1993). The source and target items are well aligned functionally when in fact is used in a context of strongly opposed propositions and serves to explicate (a surprising and unexpected) contrast like in (1). Then, na samom dele is used in its natural central meaning of refutation (Iordanskaja and Mel’cuk, 1999).

(1) EN_1_49.txt: The workings of the money system and the economy are always discussed in mysterious terms. People feel that it is something too complicated for them to understand. In fact, only falsehoods and false principles need to be discussed in mysterious terms.

However, when the proposition introduced by in fact supports the idea expressed earlier and the context lacks ‘surprising’ quality, na samom dele sounds out of place. In example (2) the importance of ‘watching advertising costs’ is supported by devoting the whole chapter to the issue, while the usage of the DM in the Russian target is incongruous with the usual contrastive reading of the sentence-initial na samom dele. Example (3) apart from demonstrating the same translational inadequacy, is an example of another source of na samom dele in Russian translations – actuality and certainty expressions.

(2) EN_1_66.txt: I believe that watching advertising costs is important. In fact, a future chapter is dedicated to getting the most from your agency at a fair price.

RU_1_66_1.txt: Я считаю, что следить за расходами на рекламу очень важно. На самом деле, следующая глава посвящена получению максимальной отдачи от...
(3) EN_1_2.txt: Diamonds are neither valuable nor rare. Though fabulously expensive, they are actually one of the most common minerals on earth. ... (supporting arguments)... In fact, without the tradition and romance which have always given diamonds their sentimental value, they would be almost worthless.

RU_1_2_16.txt: Алмазы не являются ни ценными ни редкими. Хотя они необычайно дорогие, на самом деле они являются одним из самых распространенных минералов на земле. ... На самом деле без традиционной романтики, которая всегда придавала алмазам их сентиментальную ценность, они ничего не стоят.

This group includes words and phrases like actually, indeed, really, in truth, obviously, in reality, in the real world, certainly, in effect. It is only in contexts of statement of fact as assertions of actuality where the use of the adverbial na samom dele is justified. Nonetheless, translators force the latter into contexts of confirmation, too, presumably, mixing it up with v samom dele (в самом деле) due to the strong formal resemblance of the two (Iordanskaja and Mel’cuk, 1999). The third most frequent translational pattern that contributes to the unusual frequency of na samom dele is explicitation of contrastive relations originally expressed by a conjunction or implied. Na samom dele is added by translators in 47 out of 117 occurrences and all of them have contrastive contexts like in (4). This trend is especially visible when the source clause contains just negation as an emphasized contradiction to the statement above (see (5), where the urge to use this DM overrides the considerations of style – the clumsy lexical repetition).

(4) But talk to scientists in private, and many will grumble. Но на самом деле, многие в исследовательских кругах остаются недовольны.

The analysis of professional translations is indicative of the same tendencies, even if they are milder. It is our feeling that на самом деле is currently changing its frequency distribution in Russian as a whole and as a result becomes more semantically diluted, effectively turning into a junk filler word. Whether this trend really exists and whether it is induced by out-of-English translations are matters of separate research. Generally, ‘translationally difficult’ DM, both leading to inflated frequencies of some Russian items and to violated patterns of text cohesion, can be revealed by looking at contrasts in translational strategies between learners and professionals.

References


Assessing the diachronic change of a word-formation pattern:
Frequency, productivity, and interaction patterns
Stefan Hartmann (University of Hamburg, Germany)

Nominalization with the suffix -ung (e.g. Landung ‘landing’, Bildung ‘education’) is certainly among the most well-studied word-formation patterns in German, both synchronically and diachronically (see Hartmann 2016 for a recent overview). Most importantly from a historical-linguistic perspective, Demske’s (2000) corpus-based study of ung-nominalization in the Early New High German period (1350–1650) has shown that the pattern, while becoming more frequent, suffers a considerable decrease in morphological productivity. In addition, she argues that the word-formation pattern becomes more “nominal” over time: More and more ung-nominals denote concrete objects (e.g. Heizung ‘heating device’) or even persons (Bedienung ‘waiter/waitress). While much of Demske’s (2000) study remains qualitative, Hartmann (2016) has conducted a systematic quantitative corpus study using larger corpora and extending the scope of investigation from the Early New High German period to the beginning of the New High German period (from 1650 onwards). Using the Mainz Early New High German Corpus (Kopf 2016) and the GerManC corpus (Durrell et al. 2007), it could be shown that ung-nominals, over time, are increasingly attracted to prototypically “nominal” constructions (e.g. determiner constructions and plural constructions), while their frequency in constructions that evoke a verb-like construal drops significantly.

One drawback of Hartmann’s (2016) study, however, is that the corpora are fairly small and only comparable to a limited extent. In addition, the study suggests that the relevant changes are still in progress at the end of the time period covered by the GerManC corpus, which is why it seems promising to investigate the pattern in a corpus that covers the 19th century as well. Therefore, the present study replicates the results obtained by Hartmann (2016) using the German Text Archive (Deutsches Textarchiv, DTA). The DTA is a 100-million-word corpus covering the time span from 1600 to 1900. As the DTA is very unbalanced both for time periods and for text types, a smaller subcorpus (DTAbaby) has been compiled comprising 270 texts of normalized length, balanced for fifty-year periods and three text types, and comprising about one million tokens. Fig. 1 shows an overview of the composition of both corpora.
Fig. 1: Overview of the absolute token frequencies of DTA (left panel) and DTAbaby (right panel). Note that newspaper texts have been omitted from DTAbaby as they are heavily underrepresented in the DTA data.

<table>
<thead>
<tr>
<th>Period</th>
<th>Tokens</th>
<th>Types</th>
<th>Hapax Legomena</th>
</tr>
</thead>
<tbody>
<tr>
<td>Period 1 (1600-1649)</td>
<td>922</td>
<td>335</td>
<td>84</td>
</tr>
<tr>
<td>Period 2 (1650-1699)</td>
<td>924</td>
<td>384</td>
<td>93</td>
</tr>
<tr>
<td>Period 3 (1700-1749)</td>
<td>1273</td>
<td>395</td>
<td>69</td>
</tr>
<tr>
<td>Period 4 (1750-1799)</td>
<td>2106</td>
<td>501</td>
<td>64</td>
</tr>
<tr>
<td>Period 5 (1800-1849)</td>
<td>2720</td>
<td>614</td>
<td>102</td>
</tr>
<tr>
<td>Period 6 (1850-1899)</td>
<td>3001</td>
<td>663</td>
<td>159</td>
</tr>
<tr>
<td>Sum</td>
<td>10946</td>
<td>2892</td>
<td>571</td>
</tr>
</tbody>
</table>

Tab. 1: ung-nominalization in the DTAbaby corpus: Overview of the corpus data.

Using the DTAbaby data, it can be shown that unger-nominalization experiences a steep increase in token frequency throughout the entire period covered by the corpus, as Tab. 1 shows. Using Baayen’s (e.g. 2009) measure of potential productivity, i.e. the ratio of hapax legomena to the total number of items belonging to the construction in question, suggests that the productivity of the pattern decreases until the end of the 18th century but then sees a slight increase. However, given the significant differences in token frequency, comparing the potential productivity values of the individual corpus periods is highly problematic (see e.g. Gaeta & Ricca 2006; Säily 2011, among many others). For this reason, a finite Zipf-Mandelbrot model (see e.g. Baayen 2001; Evert 2004) was used for extrapolating the number of hapaxes that can be expected for an arbitrarily large total number of tokens. As the Zipf-Mandelbrot model, unlike Baayen’s productivity measures, does not require equal sample sizes, all 1,713,147 ung-nouns attested in the complete DTA corpus were used for obtaining the extrapolations. As the goodness-of-fit of the resulting model proved suboptimal, a bootstrapping approach was used in addition to the simple model. For each of the three centuries covered by the DTA, 100,000 attestations were randomly sampled, and a Zipf-Mandelbroot model was fit to the data using zipfR (Evert & Baroni 2007). This procedure was repeated 100 times. The left panel of Fig. 2 shows the results. What appears, in the plot, as a thick dark-grey line consists of 100 individual lines that represent the FZM models fit to the 17th century data. The same goes for the (partly overlapping) areas that appear in somewhat lighter shades of grey which represent the 18th and 19th century data, respectively. The black lines represent the average growth curve, obtained by calculating the mean $V_1$ (= number of hapaxes) for each $N$ (= token frequency). The right panel shows the extrapolated potential productivity for an arbitrary value of $N=500,000$. 
In sum, the extrapolated values suggest that the morphological productivity of the pattern continues to decrease, as observed by Demske (2000) and Hartmann (2016) for other corpora. While the decrease is clearly observable from the 17th to the 18th century, the picture for the 18th/19th century is not entirely clear – while some models suggest a slight increase in potential productivity, others suggest that it keeps decreasing. Nevertheless, the overall picture is very clear and confirms the findings from previous literature that all in all, the morphological productivity of unger-nominalization wanes throughout the New High German period.

However, equally important for understanding the diachronic developments of a word-formation pattern are what can be called “interaction patterns” from a Construction Grammar perspective, according to which more abstract morphological or syntactic patterns are considered form-meaning pairs (constructions) in their own right (cf. e.g. Goldberg 2006). Demske (2000) and Hartmann (2016) have already shown that the pattern’s constructional preferences change significantly over time, reflecting its drift towards a higher degree of “nouniness”. Three constructions are particularly interesting in this regard: The determiner construction, the plural construction, and what can be called the [P NOM] construction, in which a nominalization is used as the complement of a preposition (without a determiner). The latter is particularly interesting in that it tends to evoke a highly processual construal of the nominal in question. Consider, for instance, grabung ‘digging’ in (1) (from Demske 2000: 380).

(1) Diese wochen hat man allhie in grabung deß Grunds zu S. Petro ein Kreutzlein oder heyligthum [...] gefunden. ‘This week, in digging the ground of St Peter’s [cathedral], a cross or sanctuary has been found.’ (Relation des Jahres 1609)
The proportion of *ung*-nominals in the [P NOM] construction decreases constantly throughout the period covered by the DTA*baby* corpus (Kendall’s $\tau=-1$, $T=0$, $p<0.01$; see Fig. 3).

Conversely, the proportion of *ung*-nominals with a determiner and the proportion of pluralized *ung*-nominals continue their increase that started in the Early New High German period (see e.g. Demske 2000). Both the use of determiners and pluralization can be seen as diagnostics of increasing “nouniness” (see e.g. Vogel 1996; Fonteyn 2016; Bekaert & Enghels forthc.). The proportion of *ung*-nominals with a determiner increases slightly, but significantly (Kendall’s $\tau=1$, $T=15$, $p<0.01$). In the case of pluralization, the initial increase is followed by a slight decrease, but altogether, it is fairly clear that we find more pluralized instances in later than in earlier corpus periods.

Fig. 3: Relative frequency of *ung*-nominals in [PREP NOM] constructions in relation to the total number of *ung*-nominals in the respective corpus period as attested in the DTA*baby* corpus.

Fig. 4: Relative frequency of *ung*-nominals with a determiner (left panel) and of pluralized *ung*-nominals (right panel) in relation to the total number of *ung*-nominals in the respective corpus period.
All in all, then, the data from the DTA corpus and from the newly-compiled sample corpus DTAbaby lend further support to the hypothesis that the diachrony of ung-nominalization can be described as a “nominalization process with ‘nominalization’ taken literally” (Demske 2002: 69). Such processes of “nominalization” can be observed in other cases, and in other languages, as well. For instance, Fonteyn & Hartmann (2016) have shown that English ing-nominals undergo a surprisingly similar development. But while the tendency of abstract nouns to develop more concrete meanings – posited by e.g. Panagl (1987: 146) as a widespread cross-linguistic tendency – has so far been an observation largely based on qualitative analysis of the available data, corpus-linguistic studies like the one presented here can help empirically substantiate such hypotheses drawing on in-depth quantitative analyses of individual word-formation patterns.

References

Deutsches Textarchiv (DTA). http://deutschestextarchiv.de/


Multilingualism in Chile: representations of languages in Chilean news media discourse
Hanna Sliashynskaya (Newcastle University, UK)

Linguistic ideologies, or beliefs about languages and their use, are key to dynamics and changes in language choice, language minorisation and death. Linguistic ideologies, especially those of monolingualism, have long been part of policies of nation-states (Fairclough, 2015; Shohamy, 2006) despite the prevalence of multilingualism in social domains (Meyerhoff, 2008). Chile, the context of this research project, is a multilingual country with a surprisingly limited amount of language planning and policy legislation (Leclerc, 2015). In view of such laissez-faire regulations of linguistic setting, the question arises as to what causes indigenous languages to lose speakers whilst “big” languages expand further.

Thus, this research project examines the multilingual context of Chile and how dominant and minoritised languages are represented in popular national online newspapers. The collected data includes 8877 news articles published in ten most widely-read Chilean online newspapers between 2010 and 2016 and containing references to Chile’s local (Mapudungún, Rapa Nui, Aymara, Quechua, Yámana, Huilliche, Qawasqar, Kunza and Spanish) and foreign languages (English), as well as bilingualism and multilingualism. The period was established due to rising importance of language in indigenous issues (Rojas, 2016) and an increase in foreign language educational policies (Minsegpres et al. 2014) in these recent years. A corpus of 3 717 129 words was compiled to reveal how media represent languages and what discursive strategies are used to conduct metalinguistic debate. Corpus-based critical discourse approach was used to analyse the data.

An exploratory analysis of the “aboutness” of the corpus has shown that español ¹(Spanish, freq. 4712) and inglés (English, freq. 3749) appear among the top 10 most frequent lexical words in the corpus which is indicative of their prominent presence in linguistic talk in the Chilean online newspapers. For comparison, only chino (Chinese, freq. 654) and mapudungún (freq. in all spelling variations 1197) appear among top 10 of two sub-corpora. Such frequencies suggest that the media tend to report more on dominant languages which can be explained by Spanish being the official language of the country and English the only foreign language taught in public education. However, a closer look at the context and use of further techniques is necessary to explain why this is the case (Baker, 2006, p. 56).

As this study focuses on representations, the corpus tool to approach the analysis beyond wordlist was collocation, as meanings of collocates of a word contribute to the meaning of the word itself (Baker et al., 2008, p. 278). When looking at collocates of ‘bilingüe” in its different forms (1 025 occurrences), languages appeared among significant collocates: Spanish (MI 7.61), Mapudungún (MI 6.29) and English (MI 5.88) among others.

Although the strength of collocation suggests otherwise, English is linked to bilingualism much stronger than other varieties in Chilean online news. This became apparent upon examination of concordance lines.

¹ Translations from Spanish undertaken by the author.
Some examples from the corpus show how often in reporting bilingualism is understood and reported as the ability to speak English.

This is clear from speeches of educational authorities (see Figure 1), which try to justify the necessity to learn English appealing to a ‘national goal’. Emol.cl here quotes the Minister of Education Joaquín Lavín. In his speech, he uses deictic ‘lo que queremos’ (what we want) and ‘todo el mundo’ (everyone) when justifying the reasons for learning English which hardly builds up a strong argument. From this quote, it becomes clear that Chile has not yet started to ‘become’ a bilingual nation despite numerous indigenous languages spoken in the country. However, according to the minister, it is high time that bilingualism in Chile is measured. Here a presupposition that readers understand bilingualism as the ability to speak English is present, as this piece of news reports on the introduction of a new nation-wide English test.

"... lo que queremos es que Chile comience a medirse en términos de inglés. Todo el mundo dice que Chile debe convertirse en un país bilingüe, pero en algún momento hay que partir"’... what we want is for Chile to start measuring itself in terms of English. Everyone says that Chile must become a bilingual country but at some point, we need to start’

(Emol.cl, 2010)

Figure 1 Bilingualism and English language in education debates

Another example of such assumption on behalf of the authorities is illustrated in Figure 2. In addition to reinforcing the idea of monolingual Chile (‘the intention of our country "to become a bilingual nation”’), Chancellor Moreno in his meeting with Hillary Clinton stated that to achieve a bilingual nation it is necessary to bring teachers from abroad, that is, native speakers of English. In line with the previous example, the presupposition here is that speaking English is a universal good that children should not be deprived of. Thus, this example shows once again the superior position of English in educational discourse and demonstrates the ‘myth of native speaker’ (Davies, 2003) in educational debates, or the idea of superiority of native speakers when it comes to teaching a language.

‘uno de los temas centrales fue la intención de nuestro país “de convertirse en una nación bilingüe donde todos los niños tengan oportunidad de hablar inglés”, para lo cual ambos estudiaron opciones sobre “cómo trasladar profesores de inglés”’

‘One of the central topics was the intention of our country “to become a bilingual nation where all children have the opportunity to speak English”, for which both explored the options of "how to bring in teachers of English”’

(Emol.cl, 2011)

Figure 2 Bilingualism and English in education debates
On the other hand, Mapudungún also appears in connection to bilingualism in educational debate when the possibility of opening bilingual kindergartens is discussed on BioBio.cl. However, here the discourse of endangerment and death of the language is dominant whilst no emphasis is made on individuals or the country becoming bilingual (see Figure 3).

"Así como no podemos entregar tierras sin apoyo productivo tampoco podemos dejar que de aquí al 2030 no quede ningún hablante de nuestras lenguas indígenas, porque un pueblo sin lengua tiende a desaparecer", señaló el director. (Lavado, 2012)

"Just as we cannot give away land without productive support, we cannot allow that from now until 2030 there is not a single speaker of our indigenous languages left, because a people without a language tends to disappear," said the director.

Figure 3 Discourse of endangerment in educational debates

The above examples relate to the idea of “elite” and “folk” bilingualism (de Mejía, 2002; Guerrero, 2010) when bilingualism is perceived rather as the ability to speak two European languages, whilst indigenous languages are not seen as adding up to bilingual condition but rather as “a dialect that is part of their identity and their sense of belonging” (Lavado, 2012).

These and other examples of metalanguage in the corpus demonstrate the media’s affordances to transmit and reinforce linguistic ideologies, often in very subtle ways. In view of scarce language policies in Chilean context, the possible effect of media coverage of language matters on the public opinion and perceptions of them cannot be underestimated and the media’s representations of dominant and minoritised varieties can contribute to maintaining the linguistic status quo.

References

Hillary Clinton recibe al canciller Moreno y destaca el "liderazgo de Chile en el mundo". (2011, August 3). Emol.cl Retrieved from http://www.emol.com/noticias/internacional/2011/08/03/496066/hillary-
clinton-recibe-al-canciller-moreno-y-destaca-el-liderazgo-de-chile-en-el-mundo.html


This paper has three interrelated aims. The first aim is to present a novel corpus-based methodology for the diachronic analysis of generic structure. The second aim of the paper is to present the results of an empirical study which used this methodology to identify and study changes in the generic structure of a diachronic corpus of British patent texts spanning three centuries. The third aim of the paper is to consider the implications of my diachronic analysis of patent texts for a key theoretical question in current diachronic genre studies: is genre change best understood as a Darwinian, evolutionary process, or is it better understood as a series of Kuhnian, revolutionary paradigm shifts?

The methodology to be described in the first part of this paper follows Biber and Conrad (2009) in regarding generic structure as being principally marked by overt textual features such as section headings and subheadings, fixed phrases, special formatting and so on. The first stage of the methodology therefore involves identifying these overt textual features. The function(s) of the generic stages marked out by these overt features are then established by qualitative analysis, and each stage is given a simple identifying code letter and/or number. To illustrate, Figure 1 shows a generic structure coding for patent no. 1700 (‘Elin’s Improvements to Shoe Buckles’), which was published in 1789:

<table>
<thead>
<tr>
<th>Code</th>
<th>Function of generic stage</th>
</tr>
</thead>
<tbody>
<tr>
<td>z</td>
<td>Identifying addressees of the text</td>
</tr>
<tr>
<td>p1</td>
<td>Identifying author of the text</td>
</tr>
<tr>
<td>y1</td>
<td>1st person statement that monarch has granted patent protection</td>
</tr>
<tr>
<td>x1</td>
<td>1st person statement of condition of grant</td>
</tr>
<tr>
<td>w1</td>
<td>1st person statement of fulfillment of condition of grant</td>
</tr>
<tr>
<td>v</td>
<td>Description of invention</td>
</tr>
<tr>
<td>u1</td>
<td>1st person witness and/or signature</td>
</tr>
<tr>
<td>q</td>
<td>Other witness signatures</td>
</tr>
<tr>
<td>te</td>
<td>English-language confirmation that specification has been lodged on a particular date</td>
</tr>
<tr>
<td>n</td>
<td>Drawings</td>
</tr>
</tbody>
</table>

Following this method, the generic structure of each complete exemplar text is thus reduced to and represented as a relatively simple code string, such as zp1y1x1w1vu1qte1n for the patent described above. These code strings can then be processed using adaptations of standard corpus analysis techniques. The key technique for identifying patterns of diachronic change involves using
an adaptation of dispersion plotting to identify when and how particular variant forms emerge, how long they persist for, and whether and how they mutate or disappear from use altogether over time.

In the second part of the paper I will present the results of an empirical study which used this methodology to identify and study changes in the generic structure of a corpus of British patent specification texts between 1711 (the year in which the world’s first patent specification was published) and 2011. The corpus was compiled from the complete collection of over 2 million historical UK patent documents held at The British Library, and (for more recent texts) from the European Patent Office’s Espacenet online patent search interface (http://www.epo.org/searching-for-patents/technical.espacenet.html#tab1). On the basis of this analysis I will argue that there have been five major transformations in the structural form of the British patent specification genre in its three hundred years of continuous existence. I will interpret these generic changes in social and functional terms, showing how they can be related to concurrent changes in intellectual property law and its conceptual underpinnings, to developments in science and technology, to the growth of manufacturing industry and other forms of commercial activity during the period, and to broader developments in British society and politics as a whole.

In the final section of my talk I will discuss the implications of my empirical analysis of patents for a current theoretical question in diachronic genre studies: is the process of genre change best understood in Darwinian terms, as a matter of constant and gradual evolution (e.g. Gross et al, 2002), or is it better understood in Kuhnian terms, as a series of relatively stable periods of activity punctuated by occasional and abrupt revolutionary shifts (e.g. Berkenkotter, 2009)? In general, the results of my study lean more towards a revolutionary than an evolutionary account of genre change. However, I will also point out a number of features which seem to be better described in evolutionary terms. Ultimately, I will caution against an overly literal Kuhnian interpretation of my data. I will also suggest that the aptness of either an evolutionary or a revolutionary interpretation of the results of a diachronic genre analysis may also depend to a considerable extent on the function and status of the genre in the society in which it operates, and on the level of the genre analysis itself.

References


Action, metaphor and gesture: A corpus-analytical approach
Yao Tong (Vrije Universiteit Amsterdam, the Netherlands) and Alan Cienki (Vrije Universiteit Amsterdam, the Netherlands / Moscow State Linguistic University, Russia)

We experience the world around us through actions: we reach for something, hold it, or play with it. Meanwhile, we also share our kinesthetic experiences with other people using language and gesture. Due to their fundamental role of actions in human experience, action verbs encompass a rich pool of source domains for linguistic metaphorical extensions (e.g., grasp/hold/pick up/throw away an idea). But conceptual metaphors, as shown in many studies, are not only expressed in the spoken modality but also in the manual modality, through gesture (e.g., Calbris 2003, Cienki and Müller 2008, Sweetser 1998). Particularly, claimed as simulated action (Hostetter & Alibali 2008), gesture is capable of visually representing specific source domain information (i.e., embodied aspects) of abstract meaning, which is not expressed in speech (e.g., gesture can show exactly how “limits” can be pushed).

The current research aims to investigate how our linguistic and gestural behaviors are potentially motivated by our kinesthetic experiences using a corpus-analytical approach. It examines linguistic semantic/metaphorical extensions of action verbs and their spontaneous co-verbal (referential) gestures within naturally occurring language data. English action verbs collected from the action ontology IMGACT (http://www.imagact.it/imagact/query/gallery.seam) (Moneglia et al., 2012) are analyzed in terms of their linguistic patterns: (1) frequency in large corpora (COCA and BNC); (2) semantic complexity; (3) semantic frames (e.g., deterioration/modification of the object).

Based on the linguistic behaviors of action verbs, partial sets of them which tend to be extended to more abstract meanings are examined for their gestural behaviors. Whereas co-verbal gestures depicting concrete actions can be straightforward and specific via iconic representation (e.g., gestures with pushing a button on plane), it remains unclear how specific gestures depicting metaphorical meanings usually are (e.g., gestures with pushing limits in a political debate). All the gestural data are collected from the Distributed Little Red Hen video database (https://sites.google.com/site/distributedlittleredhen/home), which allows for the search of relevant words and phrases through the closed-captioning recorded with over 250,000 American televised programs. Gestural patterns are analyzed in terms of the form features which might be able to capture a wide range of the form variation and modification, for instance, hand shapes, location in space, movement trajectories (Bressem, Ladewig, & Müller 2014).

A pilot study finds that hand action verbs (e.g., pull, push) are more frequent and semantically diverse than mouth action verbs (e.g., chew, swallow) and those for foot actions (e.g., kick). Gestural patterns of 5 manual action verbs (e.g., pull, push, lift, pick, hold) show that gestures depicting abstract manual actions are less varied and more recurrent in nature, compared to gestures depicting concrete manual actions. For instance, gestures depicting abstract manual actions are produced with less varied and more lax handshapes. We will discuss how the current
research can provide insight into how what we know about embodied cognition can explain linguistic and gestural patterns related to communication about our kinesthetic experiences.

References


A corpus-based approach to investigating twenty-first-century prescriptivism
Morana Lukac (Leiden University, The Netherlands)

1 Introduction

The publication and the remarkable success of Lynne Truss’s prescriptive guide to punctuation Eats, Shoots and Leaves (2003) led Beal to observe that prescriptivism had returned with a vengeance in the twenty-first century (2010). Whereas the number of publications of similar books containing prescriptive usage advice is continuously on the rise (Tieken-Boon van Ostade, in progress), not all linguists share Beal’s pessimistic view on the current quality of public debates on language use. Crystal, for instance, claims that at the moment we are “moving away from an institutionalized prescriptivism towards a more egalitarian linguistic era” (2006), and Cameron suggests that the web may be a platform facilitating the interaction between lay people and linguists (2012) and consequently popularizing the descriptive approach to language. Empirical research on the topic, however, is lacking, rendering it difficult to establish easily which between the opposing arguments is more accurate.

In my analysis, I explore a corpus of metalinguistic comments comprising three internet registers – blogs, Wikipedia pages and below-the-line comments of online articles – and one pre-internet register – letter-to-the-editor (LTE) sections of newspapers by way of multidimensional analysis (Biber, 1988). Wikipedia talk pages hosts discussions on grammar and usage that result in creating and editing Wikipedia entries. The participants in the remaining three genres simply comment on and express their views on usage and grammar. The analysis reveals that significant variation exists regarding how usage is discussed in the analysed registers, and it can largely be explained by the existence of different communities of practice (cf. Wenger, 1999; Eckert, 2006) in which the metalinguistic comments are embedded. Wikipedia pages on which usage and grammar are discussed, the results suggest, provides a space for forming an especially close-knit community of practice. In the second part of the analysis, I compare – by means of key semantic domain analysis – Wikipedia pages that describe problematical aspects of grammar (as an example of a type of present-day online sources) with a sub-corpus of entries compiled from a diachronic database of usage advice literature published between 1770 and 2010, called the Hyper Usage Guide of English or the HUGE database (Straaijer, 2014). The results of this analysis reveal, on the one hand, that Wikipedia entries manage to obtain a high level of objectivity and avoid prescriptive accounts due to the editors’ commitment to the community principles as opposed to traditional usage guides, which contain many ipse dixit judgments. On the other hand, the results also show that we can identify continuity throughout the history of prescriptivism – as Beal and Tieken-Boon van Ostade suggest – with regard to the topics that
are discussed in usage debates.

2 The corpus and analysis

The research initiative Bridging the Unbridgeable: linguists, prescriptivists and the general public at Leiden University has traced the history of prescriptivism in the English language with the help of the HUGE database, which comprises 77 usage guides published from 1770 onwards. Next to usage guides, the liveliest arenas for discussions on usage today can be found online. In the present study, I explore these arenas within a self-compiled corpus of metalinguistic comments comprising three internet registers – blogs, wikis and below-the-line comments of online articles – and one pre-internet register – letter-to-the-editor (LTE) sections of newspapers – in all of which members of the public express their views concerning usage.

In building the corpus of metalinguistic comments, I implemented Biber’s proposal for representativeness in corpus design (1993), in which a set of objective criteria is introduced for determining the number of texts that need to be included in corpus analysis. The underlying idea of Biber’s proposal was: the more variation, the more texts are needed to represent a genre, and the less variation, the fewer texts. Although Biber’s aim was to achieve objectivity, some amount of subjectivity cannot be avoided in setting up a study. I needed to thus arbitrarily decide on a number of texts that I would include in the final corpus: I established that the number would amount to a total of 1000 texts, which is approximately twice the size of the corpus Biber compiled in his 1988 study (57) (481 texts), since the texts in the genres I included tended to be much shorter than the ones from the original study. This procedure led me to establish my final corpus design shown in Table 1.

Table 1. The corpus of metalinguistic comments

<table>
<thead>
<tr>
<th>Genre</th>
<th>Blog</th>
<th>Wiki</th>
<th>BTL</th>
<th>LTE</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Texts / Words</td>
<td>6,441 / 107</td>
<td>15,850 / 181</td>
<td>27,343 / 238</td>
<td>131,455 / 474</td>
<td>181,089 / 1,000</td>
</tr>
</tbody>
</table>

In the next step, I allocated the four registers to Biber’s six dimensions with the help of the Multidimensional Analysis Tagger (Nini, 2015). After that, I set out to identify the extent to which the four registers were similar to each other and to the 23 registers introduced by Biber (1988) by performing a statistical analysis called hierarchical clustering; in the analysis I considered the six dimensions jointly (see Berber Sardinha, 2014). The results of the hierarchical clustering analysis are reported on in Figure 1.
Much of what we can observe from the general bipartite division in the tree diagram in Figure 1 indicates a distinction between spoken and written genres. The blog and BTL comments are clustered together being the most similar among the genres I analysed, and they are also closely linked to the interactive formal or written modes of communication. Letters to the editor, the only pre-internet genre among the four, are grouped together with discursive and informational genres, and they are also linked to other newspaper genres. The most distinct here are wikis, which belong to cluster B, and are grouped, on the lowest level in the cluster, together with spontaneous speeches and personal letters.

If anything, the results in Figure 1 indicate that the four registers, and consequently, the usage discussions found in them are heterogeneous. A closer examination reveals that the differences among them depend mostly on the existence of a close-knit community of practice and the communicative goals of their members. Wikis – websites that are created to enable collaborative projects
have stronger ties and are more practice-oriented than the other registers.

3 Comparing twenty-first-century to traditional usage advice

With the introduction of Web 2.0 the discourse on language use entered the new media enabling re-contextualization of the previously established one-way communication between the advice-seeking lay community and the advice-giving language experts. I compare these two contexts by looking at the Wikipedia entries and their secondary Wikipedia Talk pages on the topics related to usage and the entries in usage guides retrieved from the HUGE database. For this I used the web-based Wmatrix tool (Rayson, 2009) through which I identified key words, and subsequently key semantic domains in the two corpora that would help determine significant differences between the Wikipedia-type entries and those found in usage guides.

The results show that Wikipedia entries obtain a higher level of objectivity and avoid prescriptive accounts as opposed to usage guides. Wikipedia entries are a product of collaboration as opposed to usage guides, which are usually works of individual authors. Due to this fact, as well as the guidance of Wikipedia principles, balanced discussions on Talk pages and regular edits, Wikipedia entries on language use largely reflect critical, up-to-date accounts relying primarily on linguistics as a discipline and actual usage, instead of single authorities and traditional gate keepers. The actual impact of Wikipedia and other forms of social media on usage and their popularity when compared to other sources of advice on usage is yet to be explored.

References


https://sites.google.com/site/multidimensionaltagger/.


The idiom principle and translated texts
Gill Philip (University of Macerata, Italy)

Introduction

The idiom principle has been a mainstay of (monolingual) phraseological research for the past quarter century. First advanced by Sinclair (1991), it proposes that normal text is composed of intersecting pre-fabricated chunks, only occasionally necessitating a switch to allow for the insertion of terminological (ibid.) insertions before switching swiftly back again to the phrasal norm. Observable in corpus data via KWIC concordances – where minor variations can be absorbed – as well as through n-gram and cluster extraction which have more stringent requirements regarding fixity of word-form and syntactic sequencing, it is clear that the idiom principle does indeed hold sway in normal language use. But what happens when a text is not being constructed, per se, but rather reconstructed, as is the case with translated texts? Little is known about the extent to which the idiom principle survives the process of translation because until now very few scholars have investigated multiple translations, and none of those who have done so have investigated phraseological (rather than terminological) phenomena.

This contribution seeks to redress this gap in knowledge, drawing on the author’s own learner translation corpus. Compiled for use in the classroom, the data consists of the students’ translation assignments: in other words, multiple translations of selected source texts. These are subjected to linguistic analysis via concordancing software in the classroom, allowing students to view all versions – their own and those of their classmates – via projected concordances called up in real time. Concordancing multiple translations makes it possible to identify points of convergence (identical or similar translation solutions) and divergence (wide variety of translation solutions). Convergence reveals that the class has reached a tacit consensus on how best to translate an item or text fragment, with no evident problems arising. Divergence, in contrast, is indicative of a translation problem (such as misunderstanding of the source text or problems in rendering the meaning effectively in the target language). Unsurprisingly, it is on translation problems that most classroom discussion hinges; but beyond the immediate needs of the translation class, convergence too deserves close linguistic scrutiny. The dearth of research on multiple translations prevents us from assessing at present just how similar translated texts ought to be – they are, after all, reconstructions of an existing text in another language. What we can investigate, instead, are the ways in which convergence manifests itself, and how these relate to existing theoretical models of `normal’ text production.

Data and methods

The corpus used in this study consists of ten small subcorpora, each comprising between ten and fourteen translations into Italian (L1) of a 400-word English (L2) source text. Since at present no parallel concordancer can handle so many texts, AntConc (Anthony 2014) sufficed for the extraction of items to study in class. With no text alignment available, the search procedure relies on unambiguous items in the source text, e.g. proper names, dates and figures, etc., which serve as entry points into the data. Using the widest context span available, it is then possible to locate the item of interest on-screen.
The idiom principle, convergence, and divergence in multiple translations

When analysing the translations, it was found that convergence is a strongly phraseological phenomenon: not only were the translations of significant content words agreed upon, but so too was their phraseological environment. While this might be expected to some extent, given that we are dealing with translation rather than original text composition, this feature is of particular interest to linguists and translators alike. By way of background, it is worth noting that while the morphological simplicity of English lends itself well to strong phraseological patterning, highly inflected languages such as Italian are expected to perform less well: each verb has six distinct forms for each of its eight tenses, all nouns inflect for gender and number, and adjectives agree with these, having up to four possible forms each (masculine singular/plural; feminine singular/plural); the definite article has six forms including the contraction ‘l’, and prepositions can be merged with these as preposizioni articolate. Not requiring the specification of subject pronouns, word order too is often flexible: the object is often placed before rather than after intransitive verbs, and in translation, pronouns can be specified (mimicking the source text) or omitted, at will. Morphological variety and syntactic flexibility therefore reduce the likelihood of recurring patterns, particularly n-grams, being found – even in multiple translations of a single source text.

Considering this potential phraseological variability, we may well expect to find 3- and 4-grams in abundance, but few longer ones. However, the cases of phraseological convergence found in the data were not limited to shorter n-grams. Counting only those found in at least half of the translations for each source text (i.e. Range= min. 5), n-grams of five to eight consecutive words were common, and longer strings were often found. For example, in one of the texts, a 400-word extract from Packard (1957), there were twenty-seven 8-grams, seventeen 9-grams, ten 10-grams, five 11-grams, two 12-grams, and one 13-gram. Only one of the 8-grams occurred in all ten translations (compared to twenty-five 3-grams, thirteen 4-grams, eight 5-grams, five 6-grams, and three 7-grams), but this still leaves the fact that between five and nine independently-worked translations had many long clusters in common, providing evidence of common phraseological usage in these students’ use of their native tongue. The causes of this are yet to be unravelled: source text influence is likely to be a factor, but not the only one.

Extending the notion of phraseological regularity to include minor variation, convergence in translation becomes even more marked. It emerges clearly when viewing the translations on-screen via a concordance, and its pervasiveness confirms that these student translators largely agree on how to translate their texts. There may be minor differences in word choice, or verb tense, differences which in turn impose morphological or grammatical limitations on the immediate textual environment, but this is visibly different from divergent translations which conform to the features associated with the open choice principle (Sinclair, 1991): there is a sudden switch from the phraseological flow of convergent translations to divergence; and the switch back again is just as sudden.

Figure 1 shows an example from Packard (1957), which demonstrates first of all moderate convergence (a variety of forms of the verbs usare and utilizzare, to translate use), followed by clear divergence (the translation of tossed salad), returning again to convergence at the end of the fragment. Tossed salad – not in itself a difficult term – is evidently difficult to translate in this text: is not translated at all in translations 7-10, indicating that students were unable to interpret its meaning in context. Translations 1 and 2 simply maintain the original. Elsewhere we can see indicators of uncertainty: queries in translations 3 and 4, scare quotes in translation 7. Translations 4, 5, and 6 modify the original item by explicitation (Baker, 1993) and the student responsible for
translation 6 further explicates it (erroneously) as being metaphorical. Yet despite the evident difficulty that this item posed, the phraseology returns to normal immediately afterwards, embodying the dramatic switch between open choice and idiom principles that Sinclair (1991: 110) hypothesized.

<table>
<thead>
<tr>
<th>Translation</th>
<th>Original</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. ...utilizzò l’insalata come indicatore più affidabile...</td>
<td>Italian</td>
<td>as a more reliable indicator...</td>
</tr>
<tr>
<td>2. ...ha usato l’insalata come un più affidabile indicatore...</td>
<td>Italian</td>
<td>as a more reliable indicator...</td>
</tr>
<tr>
<td>3. ...ha utilizzato l’insalata (???) come un indicatore più affidabile...</td>
<td>Italian</td>
<td>as a more reliable indicator...</td>
</tr>
<tr>
<td>4. ...utilizzò l’insalata mista (?) come indicatore ben più attendibile</td>
<td>Italian</td>
<td>as a more reliable indicator...</td>
</tr>
<tr>
<td>5. ...utilizzò l’insalata già pronta come un indicatore più affidabile...</td>
<td>Italian</td>
<td>as a more reliable indicator...</td>
</tr>
<tr>
<td>6. ...utilizzò la metafora dell’insalatona come indicatore più attendibile...</td>
<td>Italian</td>
<td>as a more reliable indicator...</td>
</tr>
<tr>
<td>7. ...ha usato la “tossed salad” come indicatore più affidabile...</td>
<td>Italian</td>
<td>as a more reliable indicator...</td>
</tr>
<tr>
<td>8. ...usò [tossed salad] come un indicatore più affidabile...</td>
<td>Italian</td>
<td>as a more reliable indicator...</td>
</tr>
<tr>
<td>9. ...usa [the tossed salad] as a more reliable indicator...</td>
<td>Italian</td>
<td>as a more reliable indicator...</td>
</tr>
<tr>
<td>10. ...used the tossed salad as a more reliable indicator...</td>
<td>Italian</td>
<td>as a more reliable indicator...</td>
</tr>
</tbody>
</table>

Figure 1. Translations of “...used the tossed salad as a more reliable indicator of...”

Concluding remarks

This contribution will further examine cases of idiom and open choice principle in the author’s multiple variant learner translation corpus, drawing attention to the role of the idiom principle in translation, not only as product, but also as a text (re-)creating process.

References

The aim of this paper is to critically examine the combination of corpus linguistics (CL) and Critical Discourse Studies (CDS). It will do so from two angles: (i) an epistemological angle, discussing questions of knowledge generation and (ii) a social and interdisciplinary angle, debating the different questions, mindsets and skills that different kinds of researchers bring to the task, as well as the potential, often underexplored, for interdisciplinary cooperation.

The epistemological angle

Although the approach combining CL and CDS is now well established, with a twenty-year history under its belt (Hardt-Mautner, 1995; Baker, 2006), it still requires careful and indeed ongoing epistemological scrutiny. For it is all too easy to be lured into a false sense of security about the nature of the evidence merely because the analysis is computer-based. To avoid rash interpretative shortcuts, the following observations ought to be borne in mind (see Mautner, 2016).

1. Evidence that corpus-linguistic software produces is rarely self-explanatory. The program may be able to identify patterns, but it falls to the analyst to spot interesting outliers, and relate both the expected and the unexpected to the context in which the texts concerned were produced and received.

2. The presence of certain linguistic forms in a text does not lead straight to its social meaning, function and significance, and we should not pretend that it does. Although this is a truism barely worth reiterating, it is still quite common for researchers to read too much into their corpus evidence. There is a certain irony in this, because 'overinterpretation' is one of the fallacies that the use of CL is supposed to help guard against (O'Halloran & Coffin, 2004).

3. However sophisticated, every analytical tool is very good at some things but not so good at others. Dutifully acknowledging a tool’s limitations is not only a matter of scholarly integrity, but also an important step towards finding ways in which complementary tools may help compensate for any deficiencies.

4. Every corpus is an artefact constructed for a particular purpose, and caution is required when generalizing from it. As McEnery and Hardie point out (2012, p. 26), 'we must not confuse corpus data with language itself'.

5. Furthermore, there is a tendency for computer-based evidence to be rated more highly than data gathered and processed manually. This judgement may be justified in certain circumstances but misguided in others. It is most certainly counterproductive if one aims at triangulation and a rich mixed-methods approach.

6. Quantitative statements about the linguistic behaviour of an item only make sense in comparison with its use in a different corpus.

7. Both quantitative and qualitative statements about the data have their uses, and the analysis will be all the richer if both can be made. But the boundaries between them should be clear at all times, and vague quantifiers (e.g., a few, some, many, a lot) avoided.
The social and interdisciplinary angle

If epistemological critique helps us understand how knowledge is generated, it is equally worthwhile to reflect on who the key actors are in this process, and where they are positioned in the disciplinary landscape. There are two perspectives that can be usefully applied, within linguistics and outside it.

Within linguistics, the chasm appears to be widening between those researchers who are able and willing to develop new software and those who are not. The latter group is keen to have powerful analytical tools, but the reason why they want them is because they are tackling a research question at the interface of discourse and society. On the other hand, those who concentrate on the computational side of things often seem oblivious to the socio-political implications of their work. If there are calls for discourse analysts to learn how to write code and more generally to improve the rigour of their research, there should at least be corresponding calls for computational linguists to focus more on relevance, and engage more deeply in the social functions of discourse.

Finally, we should not forget that that there is a large and prolific community of discourse scholars outside linguistics, in areas such as organizational theory and behaviour for example. In those communities, corpus-based approaches are still virtually unknown. They use software for qualitative data analysis, such as NVivo or ATLAS.ti, rather than concordancers. Their hero is Foucault, not Sinclair, and they are generally concerned much more with discursive macro-structures than with linguistic detail. It should benefit both sides of the disciplinary divide to initiate a focused and productive conversation between them.

References


Is a CADS approach useful to empirically measure populism? A comparative analysis of the representation of the French “peuple” and their threats in the discourse of the major political parties in France

Dámaso Izquierdo Alegría (University of Navarra, Spain)

The notion of populism has faced different theoretical and methodological problems in political studies. Some authors define populism as a relatively stable ideology (see Mudde 2007, 2009; Stanley 2008), whereas others conceive it as a political style or strategy more or less frequently adopted by some politicians (Weyland 2001; de la Torre 2010; Moffitt & Torney 2014). Since these main views have been both widely supported and criticised, measuring how populist a politician or a political party is has become one of the central challenges in the specialized literature.

There is some consensus as for the main general features of populism, which Moffitt (2016) summarizes as (1) an appeal to ‘the people’ as opposed to ‘the (corrupt) elite’, (2) bad manners, i.e., the adoption of a provocative and politically incorrect style, and (3) a marked insistence on the idea of ‘crisis’. However, many methodological difficulties have been identified for objectively identifying and measuring populism beyond prior assumptions as to which political parties and leaders are ‘populist’ or ‘mainstream’.

In recent years, discourse has played a major role in empirically evaluating how populist politicians and political parties are. Most of these discursive approaches to populism follow manual content analysis methods: some are based on rubrics designed to assign grades to texts according to different indicators associated to populism (see, for example, Hawkins 2009, 2010; Bruhn 2012), and others focus on the proportion of text devoted to topics and notions traditionally related to populism (Jagers & Walgrave 2007; Reungoat 2011). There have also been some incursions into computer-assisted methods (Armony & Armony 2005; Pauwels 2011; Rooduijn & Pauwels 2011; Rooduijn 2013), but they have been predominantly carried out by specialists in political studies with little knowledge of Corpus Linguistics, and they rely almost exclusively on very simple collocations or on the frequencies of a priori lists of keywords related to populism.

In this paper we advocate more complex computer-based analyses following the methodology of Corpus-Assisted Discourse Studies (CADS) as a productive approach to more accurately measure to what extent a politician and his/her discourse are populist with the help of highly valuable and significant quantitative and qualitative data.

This is done here through a case study. We have analyzed how the ‘people’ and the threats they are reportedly facing are represented in the discourse of the three current major political parties in France: two ‘mainstream’ (Parti Socialiste and Les Républicains, named until 2015 Union pour un Mouvement Populaire or UMP) and one usually labelled as ‘populist’ (Front National). The CADS approach has proved to be very productive to show how different social groups, institutions, events and other entities are
represented in many types of discourse: see, for example, the analysis of the representation of Muslims (Baker, Gabrielatos & McEnery 2013), asylum seekers (Gabrielatos & Baker 2008), migrants (Taylor 2014), the European Union (Marchi & Taylor 2009) or the Iraq war (Morley & Barley ed. 2009) in the media that have been carried out with a CADS perspective, to name a few.

We have analyzed three comparable corpora comprising press releases, speeches and other relevant texts published in the News section (“Actualités”) in the website of the three political parties during one year (September 2015–August 2016). This study focuses on various types of collocations related to the words used to refer to the people and their potential enemies and threats, in line with the queries made in prominent CADS publications.

Since a constant appeal to the people and emphasis on the idea of crisis are two widely accepted distinctive features of populism, very significant differences were expected to be found between the results obtained in the Front National corpus, on the one hand, and the Parti Socialiste and Les Républicains corpora, on the other hand. More specifically, we expected that collocations would uncover an overrepresentation of the French people and some particular groups as directly or indirectly threatened by the elite in the Front National discourse. Although this hypothesis is overall confirmed by the quantitative and qualitative data obtained, some queries show that also in the discourse of the so-called French ‘mainstream’ parties (Parti Socialiste and Les Républicains) the people is more often than expected portrayed in a ‘populist’ way, similarly to the Front National discourse. This finding seems to favour the conception of populism as a political style rather than a stable ideology. We conclude with some methodological considerations concerning the validity of the data and the orientation of future comparative research to eventually develop reliable automatic tools to detect and evaluate populist texts based on significant empirical data.

References


Exploring Corpus Data Using Data Mining: A Project of Automatic Term Recognition
Dominika Kováříková (Institute of the Czech National Corpus, Charles University in Prague, Czech Republic)

The corpus data are so vast and rich that they cannot be manually processed by a person. For a corpus-based research, we need concordancers and other textual analysis tools. Data mining is the next level of corpus data processing. It is a (semi)automatic process of discovering meaningful patterns in substantial quantities of data using computer algorithms (Witten and Frank 2005, 5). Data mining is very useful for a corpus-based exploration of large amounts of texts, especially when enriched by details behind the concordances: word frequencies and distribution, contextual information, or linguistic information about each of the tokens.

Data mining can be used in various linguistic research tasks such as text classification (Sebastiani 2002), topic identification (Clifton et al. 2004), study of linguistic properties of different types of texts (i.e. scientific texts, Teich and Fankhauser 2010) or examination of human processing of lexicon (Baayen 2005). It is also able to automatically identify specific types of lexical items such as terms (Kováříková 2014), collocations or idioms.

The presented research focuses on automatic term recognition using data mining. Thanks to the opportunity to work with very large and diverse data of the Czech National Corpus (100 million words in an easily accessible corpus SYN2010 that includes academic texts of almost 40 disciplines), and thanks to the chosen approach of data mining, the scope of the research was quite wide: single-word terms as well as multi-word terms were identified in a number of academic disciplines. The research provides not only a list of automatically identified terms in Czech academic texts but also valuable findings about terms and terminology in general. In addition, the disciplines can now be compared based on the retrieved terms (number of terms, lists of the most frequent terms, terms shared by two or more disciplines).

For the project of automatic term identification, the data mining platform WEKA (Witten and Eibe 2005) required data in form of a very large matrix with thousands of lines and tens of columns. Each line was assigned to one token/text position (or a bigram for multi-word term identification), whereas each column contained values of one of the characteristics relevant for recognizing terms in a text. The features suitable for term identification are usually linguistic (e.g. POS, word structure), or are based on word frequency and distribution in texts (e.g. frequency in a given academic discipline, distribution in disciplines), or they monitor the context of the word (e.g. average distance of two words in texts). Data mining is able to track complex non-linear relations between individual term characteristics which results in effective term identification.

In addition, data mining has the capacity to identify the most relevant of the examined features (feature ranking). The highest-ranked characteristics of single word terms are derived from distribution and the frequency of a given lexical item:

1. the ratio of the relative frequency in a given academic discipline to the relative frequency in the reference corpus (containing fiction and journalism);
2. average reduced frequency (shows evenness of distribution throughout the corpus as well as the frequency of the word in corpus, see Savický and Hlaváčová 2003);
3. relative distribution in academic disciplines;
4. standard deviation of the relative distance of two neighboring occurrences of the word.

Based on the highest-ranked features, a term can be simply described as a lexical item specific (and often unique) to a particular discipline. Such description of a term can complement standard term definitions such as ISO 1087-1:2000: Term is a „verbal designation of a general concept in a specific subject field."

The features evaluated as the most important for multi-word term recognition\(^1\) are:
1. lexical association measure t-score;
2. the first word in bigram is a term;
3. the second word in bigram is a term;
4. lexical association measure mutual information (MI-score).

A multi-word term can be described as a multi-word lexical item (typically a collocation), usually with at least one of its components being a single-word term.

The data mining process was evaluated by standard statistical measures, i.e. accuracy, precision and recall\(^2\) (Manning and Schütze 2000). The results are quite promising (see Table 1) – the high success rates guarantee the reliability of findings that are based on automatically identified terms, such as the estimation of percentage of terms in academic texts or the analysis of relations between academic disciplines based on shared terms.

<table>
<thead>
<tr>
<th></th>
<th>Single-word terms</th>
<th>Multi-word terms (bigrams)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accuracy</strong></td>
<td>94%</td>
<td>97%</td>
</tr>
<tr>
<td><strong>Precision</strong></td>
<td>85%</td>
<td>81%</td>
</tr>
<tr>
<td><strong>Recall</strong></td>
<td>72%</td>
<td>75%</td>
</tr>
</tbody>
</table>

Table 1. Results of the term recognition method using data mining

The estimate percentage of text positions occupied by terms differs for individual academic disciplines, ranging from less than 10% in psychology and philosophy to almost 35% in botany and geography. In humanities, the texts contain considerably less terms (app. 17%) than in natural and formal sciences (app. 27%). The average number of terms in academic texts is app. 22%.

Another way to compare disciplines, especially in terms of finding similarities and relations between them, is to examine shared terms – in this case a great number of automatically identified terms in various academic fields. Disciplines sharing terms is a common phenomenon. There are several main reasons why a term occurs in more than one discipline\(^3\):
1. There is a relationship of some kind between disciplines, such as in case of biology and medicine, or engineering and computer science.

---

1 Data are in the form of bigrams.
2 Accuracy is a statistical measure that is able to assess the proportion of the correctly labeled words in the text (terms and non-terms). One hundred percent accuracy means that all the true terms were classified as terms and all the true non-terms were classified as non-terms. Precision is the ratio of the correctly identified terms to all words labeled as terms (correctly and incorrectly). Recall is the ratio of the correctly identified terms to all true terms (labeled as terms and as non-terms).
3 This interesting topic is of course more complex and would deserve further research.
2. The term is used metaphorically or with different meaning in one or more areas of study (e.g. *valency* in chemistry and in linguistics, *communication* in linguistics and transportation engineering).

3. A term in one field is a non-term in other disciplines (e.g. *author* in literature and in medicine).

The disciplines that shared the highest number of automatically identified terms were engineering and computer science, law and economics, sociology and history, and chemistry and biology. However, even quite distant fields such as transportation engineering and biology shared a small number of terms, e.g. *reflexní* (reflective and reflexive), *signály* (signals), *transport* (transport), *ventilace* (vent and ventilation). Although shared terms have usually rather high frequencies in texts and are interesting from linguistic point of view, it is important to understand that the majority of the terms are unique to one discipline⁴ (or to a cluster of closely related disciplines such as chemistry and biochemistry, biology and medicine etc.), e.g. *N-nitrosofenylhydroxylamin*, *regurgitace triskupidální chlopně* (triscupid valve regurgiation).

**Conclusion**

It is evident that exploring corpus using data mining can be fruitful and valuable. Corpus data are a great match for data mining because they are very large and diverse in terms of usable features. On the other hand, data mining can provide not only a high success rate or a list of identified items (in this case terms), but also deeper insight into the linguistic theory and characteristics of the analysed linguistic phenomena.

**References**


---

⁴ Particularly due to the fact that the majority of terms are multi-word lexical items.


Cross-cultural and paradigmatic influences on lexical bundles in English academic writing
Duygu Candarli (University of Manchester, UK)

Lexical bundles, which are defined as “the most frequently recurring sequence of words” (Biber et al., 1999, p. 90), play a central role and fulfil a wide range of important functions in English academic writing. The main discourse functions of lexical bundles can be summarised as follows (Biber, Conrad, & Cortes, 2004): they present arguments and frame attributes through referential expressions (e.g. ‘in the context of’), organise texts through discourse organisers (e.g. ‘on the other hand’), and convey attitudes, certainty and hedging through stance expressions (e.g. ‘it is possible to’). Lexical bundles reflect how academic writing is framed by a particular discourse community, reveal disciplinary membership and enhance effective communication in academic writing (Cortes, 2004; Hyland, 2008). Uncovering preferred ways of organisation and meaning-making, lexical bundles can also shed light on cultural preferences in academic writing (Hyland, 2008).

A large and growing body of literature has investigated lexical bundles in published English academic writing (e.g. Ädel & Erman, 2012; Biber, 2009; Cortes, 2004; Hyland, 2008). Most of these studies focused on four-word lexical bundles and investigated their discourse functions in academic prose. The general trend that emerges from the study of lexical bundles is reliance on noun and prepositional phrases which primarily serve as referential expressions in published English academic writing. In a recent cross-cultural study, Pan, Reppen and Biber (2016) examined lexical bundles in telecommunication research articles written in English by L1 English and L1 Chinese academic writers and found that lexical bundles in the two groups differed considerably in terms of functional and structural categories. Little attention has been paid to the use of lexical bundles in research articles written in English by L1 Turkish academic writers and paradigmatic influences in research on lexical bundles. A research paradigm can be defined as “a shared belief system or set of principles on what problems are to be investigated and how to investigate them” (Cohen, Manion, & Morrison, 2013, p. 13). While there is a multiplicity of research paradigms in education research, quantitative and qualitative paradigms are considered the two major paradigms in the discipline of education (Cohen et al., 2013; Creswell, 2009). Quantitative and qualitative paradigms are held to govern the methodological procedures and influence the discoursal and rhetorical conventions of research output (Cohen et al., 2013; Hu & Cao, 2015). Hu and Cao (2015) found that the quantitative research articles in the disciplines of applied linguistics, psychology, and education made more frequent use of hedges than the qualitative research articles. Based on this finding, it was hypothesised that there would be cross-cultural and cross-paradigmatic influences on the frequencies of stance expressions and discourse organisers in English academic writing.

The aim of the present study is to identify the discourse functions of lexical bundles and investigate cross-cultural and cross-paradigmatic influences on the frequency of each functional category of lexical bundles in English academic writing in the discipline of education. The two specific research questions addressed are as follows:
1. To what extent, if any, is there cross-cultural variation in the discourse functions of lexical bundles between English research articles published in international journals and English research articles written by L1 Turkish academic writers?
2. To what extent, if any, is there paradigmatic difference in the discourse functions of lexical bundles between the quantitative and qualitative research articles in the discipline of education?

Efforts were made to construct corpora which are similar in terms of genre, size, subject matter, and date. Accordingly, the empirical research articles that were published in SSCI-indexed journals in the discipline of education were extracted from the period of 2011-2015 to compile research articles that were published in international English-medium journals and those that were written in English by L1 Turkish academic writers. At the next stage, the research articles were coded into research methodology paradigms, as in Hu and Cao’s (2015) study, according to Creswell’s (2009) descriptions of quantitative, qualitative and mixed methods research paradigms in education research. When the articles did not fit neatly into these three categories, they were coded as other. Finally, a balanced number of quantitative and qualitative research articles were selected to include in the four comparable corpora (c. 1 million each).

In order to extract lexical bundles, the frequency cut-off point was set to 20 times per million words for four-word lexical bundles, and the cut-off point for range was kept at five percent of the texts, i.e. at least in seven texts in each corpus. Lexical bundles were extracted by using WordSmith Tools Version 6.0. Four-word lexical bundles were examined in terms of their discourse functions. After coding all the discourse functions of lexical bundles, a script in the Python language was written to record the token frequencies of each discoursal category of lexical bundles (i.e. referential expressions, discourse organisers, and stance expressions) for each file, which allowed me to use inferential statistics in this study. A series of two-way ANOVAs were performed in order to compare cross-cultural (English in international journals vs L2 English of L1 Turkish academic writers) and paradigmatic (quantitative vs qualitative) influences on the normalised frequencies (per 1000 words per text) of referential expressions, discourse organisers, and stance expressions. White’s adjustment, which provides a correction for the heterogeneity of variance was used for two-way ANOVAs in R (R Core Team, 2016). When a statistically significant interaction was found, simple effects analysis was conducted in order to describe the nature of the interaction. When a significant main effect of cross-cultural influence was detected, the post hoc Games-Howell test was conducted to make further comparisons.

The results revealed a significant interaction between cross-culturality and paradigm in the frequencies of discourse organisers and stance expressions, and significant effects of cross-culturality and paradigm in the frequencies of referential expressions. Paradigmatic influences were greatest for stance expressions in that the quantitative research articles published in international journals contained more stance expressions than the qualitative research articles. On the other hand, no statistically significant difference in the frequency of stance expressions was found between the quantitative research articles and qualitative research articles written in English by L1 Turkish academic writers. An opposite trend emerged for the frequencies of discourse organisers in that the quantitative research articles written in English by L1 Turkish academic writers included significantly more discourse organisers than the qualitative
research articles written in English by L1 Turkish academic writers. No statistically significant difference in the frequencies of discourse organisers was found between the quantitative and qualitative research articles published in international journals. The key qualitative similarities and differences in the use of lexical bundles across four corpora will also be discussed. The results of this study have important implications for academic writing practices and teaching academic writing in English.

References


A cookbook of co-occurrence comparison techniques and how they relate to the subtleties in your research question
Viola Wiegand (University of Birmingham, UK), Anthony Hennessey (University of Nottingham, UK), Christopher R. Tench (University of Nottingham, UK) and Michaela Mahlberg (University of Birmingham, UK)

Introduction

The concept of 'collocation' is one of the most fundamental in corpus linguistics. Its centrality is reflected in the range of definitions and approaches that have been developed in corpus linguistics over the past decades (for detailed accounts see e.g., Evert, 2008; Gries, 2013; McEnery & Hardie, 2012). The observation and quantification of collocations has been crucially used for disambiguating different senses of words, for example, in the COBUILD dictionary project (Sinclair, 1987). Beyond the application in lexicography, the description of collocations is also relevant, among others, in the contexts of learner corpus research and discourse analysis. Although corpus linguistic approaches are inherently comparative, only a few studies have taken the analysis and interpretation of collocation beyond a single corpus. For example, Gabriellatos and Baker (2008: 11) identify 'consistent collocates' across annual newspaper subcorpora describing refugees and asylum seekers. Such studies illustrate the usefulness of collocation comparisons between corpora and highlight the need for tailored statistical approaches.

In this paper, we will discuss a range of applications of a co-occurrence comparison method that we proposed to assess collocational behaviour across corpora (Wiegand et al., 2016). The aim of the present paper is to show how variations on the method can be applied to different linguistic questions that can be answered through the comparison of collocation. In various case studies we consider the influence of factors such as narration style, editorial influence and the date of publication; in each case we discuss how these factors can be explicitly considered or removed by using different variations of the method. Through the focus on such variations we illustrate the relevance of the technique to a wide range of research questions. While the paper focuses on the applicability and implications of the comparison of co-occurrences, we will provide pointers to the relevant underlying statistical theory, which is covered in detail in Wiegand et al. (2016). These methods borrow heavily from the discipline of meta-analysis as used in medical research (see e.g., Cooper, Hedges & Valentine, 2009). However, the application and interpretation of the methods are different when considered in the context of corpus linguistics. We will highlight these differences in the paper.

Method and corpora

Our CorporaCoCo method is a systematic statistical method to analyse the differences in the collocational behaviour of words across corpora (Wiegand et al., 2016). In this paper, we briefly explain the overall approach and highlight crucial statistical concepts, but the main focus is on the application and interpretation of the method and its variations. The CorporaCoCo method directly compares co-
occurrence counts for a set of node terms with all other words in the corpora. Fisher's exact test is used with a multiple test correction. The visualisations show an effect size with an associated confidence interval providing a clear overview of the results. Both the method and visualisation tools are made available in the CorporaCoCo R package (Hennessey et al., 2017), which we will also briefly introduce. Our approach to collocation comparison is applicable to any register and can be used for the conceptualisation of discourse more widely. For illustration, the paper uses case studies comparing novels by Charles Dickens to a set of other 19th century novels.

**A simple comparison of collocation behaviour between corpora**

We compare all 15 Dickens novels to a set of other 19th century novels. We are not interested in any corpus-internal variation but effectively 'average' the contents of each corpus. The question we are asking in our case study is: "what are the differences in the lexical patterns of body part nouns in Dickens's novels compared to the set of other 19th century novels?" This question is based on the importance of body part language in Dickens's fiction as described by Mahlberg (2013). In particular, we are looking for differences between the corpora and we are specifically not considering variation within each of the two corpora. Statisticians refer to this type of analysis as a fixed-effects model. A fixed-effects model is usually applied when an effect is expected to be uniform across the data. Here, we use the CorporaCoCo method specifically to take advantage of the property of the fixed-effects model to average variation within the corpora.

**A comparison of collocations between corpora that considers the dispersion of the effect**

If we want to acknowledge that there may be a difference in co-occurrence between the subcorpora (i.e. corpus-internal variation), a random-effects model can be used. The use of random-effects models has been put forward by some corpus linguists (e.g. Gries, 2015). Our use of random-effects models represents the lack of knowledge about the difference in co-occurrence between subcorpora as some kind of random effect. Crucially, the corpus-internal variation is no longer ignored or averaged out, and one of the outputs of the method is a measure of the dispersion of the effect across the subcorpora.

The level at which we consider partitioning into subcorpora, and so for which we can identify dispersion, is important and depends on the research question. In our case study we partition Dickens's novels into five subcorpora based on the periodicals in which they were originally serialised to examine editorial impact on the dispersion of co-occurrence patterns.

**Conclusion**

This paper extends our argumentation for comparing collocation across corpora, developed in Wiegand et al. (2016), to a variety of corpus linguistic applications using fixed- and random-effects models. Based on case studies comparing Dickens's novels against other 19th century novels we have shown that a simple fixed-effects
model, like the one distinguishing narration styles, can be very powerful. We also offer techniques applying the more complex random-effects models that allow us to consider dispersion of the co-occurrence effects, for example across different periodicals, whilst only making weak assumptions about the variation across subcorpora. While we use case studies to provide concrete examples of how our CorporaCoCo method can help to address specific research questions, the paper overall aims to make a convincing case for the need and the applicability of our proposed method.

References


Gries, S. T. (2013). 50-something years of work on collocations: what is or should be next... *International Journal of Corpus Linguistics, 18*(1), 137–166.


Data Structure and Searchability in a Corpus of Automatically Detected Metaphors
Oana David (University of California, Merced, US)

1. Introduction

Since its inception, the MetaNet Metaphor Identification System has yielded hundreds of thousands of automatically detected linguistic metaphors (Stickles, David, Dodge & Hong 2016). The current paper deals with the organization and searchability of the identified conceptual metaphors, linguistic metaphors, and the sentences and metadata accompanying them. The aim is to make this valuable data available to researchers, and organize it in ways that can be searched along multiple dimensions. We present a pilot method of data transfer from the metaphor identification output, stored locally on our servers, to a publicly viewable web-based Drupal database.

The metaphoric data, extracted from a variety of corpora (including GLoWbE, English (Parker et al. 2011) and Spanish (Mendonça et al. 2011) Gigaword, BNC, and manually constructed web corpora), is incredibly diverse across several target domains. These include the domains of poverty, taxation, cancer, democracy, and the gun debate, and does so for several languages (examples be discussed here are English and Spanish). The system is designed to identify source domain and target domain evoking words in a limited set of grammatical patterns, and to match these source-target word combinations (linguistic metaphors, LMs) to conceptual metaphors that are part of the metaphor repository. For example, if identified as metaphoric, abyss of poverty would be detected as a noun-of-noun grammatical construction, and would be matched with the most appropriate candidate conceptual metaphors (CMs), POVERTY IS A LOW LOCATION and POVERTY IS A MOTION IMPEDIMENT. In this way, metaphors are not word-specific, such as would be the case in statistical and seed-based learning systems commonly used in metaphor detection (e.g., Neuman et al. 2013). This reduces the number of conceptual metaphors, resulting in an economical metaphor repository wherein lemmas are mapped to frames (domains), and frames are mapped to metaphors.
Figure 1 shows the metaphor identification pipeline, which has three main components: the repository (containing the conceptual metaphors and frames), the metaphor identification component that runs over external textual sources, and the annotation database into which the extracted sentences are deposited. Each component is further linked to other data analysis and visualization programs that aid in the processing of large amounts of data. The current proposal deals only with the treatment of the annotation database portion of this pipeline.

Every time the metaphor identification is run, it produces tens or hundreds of thousands of lines of data and deposits them into the local annotation database. Each line represents one LM in its sentential context, and is tagged with a variety of metadata about the extraction. One important piece of information is the metaphoricity score: LMs are not tagged as either metaphoric or non-metaphoric, rather as more-or-less likely to be metaphoric on a 0-1 scale (Hong 2016). Other crucial pieces of information tagged on each line of data include (but are not limited to): language, the source corpus ID, sentence ID, source domain name, target domain name, and lexical items or lexical collocations on the basis of which the detection was made. Table 1 gives an example of extracted sentences for the Poverty target domain for English and Spanish.
<table>
<thead>
<tr>
<th>Target lemma</th>
<th>Source lemma</th>
<th>Score</th>
<th>Language</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poverty</td>
<td>poverty</td>
<td>0.7550</td>
<td>English</td>
<td>His policies are vague, based on a catch-all fight against poverty, but he uses top advisors, including MCP UDF vice-president Aleke Banda, another MCP defector.</td>
</tr>
<tr>
<td>Poverty</td>
<td>poverty</td>
<td>0.928</td>
<td>Spanish</td>
<td>Madonna ayuda a David a escapar de la pobreza y de una posible muerte por enfermedad y esta gente quiere echarlo todo a perder”, añadió Banda.</td>
</tr>
<tr>
<td>Poverty</td>
<td>poverty</td>
<td>0.9265</td>
<td>English</td>
<td>With Hamas and the West locked in a standoff, the Palestinians are hurtling toward an abyss of poverty and chaos, and there seems to be no way to pull back.</td>
</tr>
<tr>
<td>Poverty</td>
<td>poverty</td>
<td>0.950</td>
<td>Spanish</td>
<td>A los ocho meses de haber asumido la presidencia de Venezuela, país rico en petróleo pero afligido por la pobreza, Chávez dijo que se siente orgulloso de lo que ha conseguido en beneficio de los pobres.</td>
</tr>
<tr>
<td>Poverty</td>
<td>poverty</td>
<td>0.8</td>
<td>Spanish</td>
<td>A los ocho meses de haber asumido la presidencia de Venezuela, país rico en petróleo pero afligido por la pobreza, Chávez dijo que se siente orgulloso de lo que ha conseguido en beneficio de los pobres.</td>
</tr>
</tbody>
</table>
2. The problem

Currently, the extraction output data is stored internally on local servers, and can only be viewed and searched with special permission, and by interfacing directly with the Unix command line. Alternatively, it can also be downloaded and circulated as rather large csv spreadsheets, but only on a one-by-one basis. Other components of the pipeline, such as the input frames, metaphors and constructions are stored and interacted with in ways different from the output data, and from each other. For example, grammatical constructions are stored as SPARQL queries and are not easily searchable, and metaphors and frames are stored in a triplestore using an OWL-defined ontology (Dodge, Hong & Stickles 2015) and are searchable via a Wiki interface. This variability across both configuration and searching of data results in a system that is not easy to navigate as a metaphor researcher.

Additionally, the output data is not compiled in a single location. Each individual extraction run is defined according to four variables: language (e.g. English), target domain (e.g. Poverty), corpus (e.g. Gigaword) and date stamp. That is to say, one data file corresponding to one extraction run has unique values for all of these four variables. This means that, if one wants to search for metaphor patterns across languages, target domains, corpora, and extraction runs, one requires some way to compile the data into one location, with similar parameters aligned for easy filtering.

3. Proposed solution

The description above shows that there are multiple constraints on data conversion, server processing speed, memory allocation, security, etc. that are placed on the metaphor identification process itself. But these need not become the problem for the data review and data searching portion of the pipeline. For this reason, we propose the migration of the data to a web-based Drupal system, wherein each of the myriad types of data output types (which require myriad viewing means in the source system) converge into one cohesive and user-friendly interface. Drupal (https://www.drupal.org/node/265726) is an open-source content management network that uses PHP data objects to create an SQL abstraction layer. This means that users can create SQL queries without the programming knowledgeability required to write SQL; it is thus a very user-friendly system for both data inputting and data viewing. Figure 2 shows a screenshot of a sample query over our pilot data uploaded into our Drupal database.
The data from the MetaNet sequential extraction runs is saved as csv files, which are then uploaded to Drupal using a mapping function that relates data categories from the csv files to data categories already entered in the Drupal database. We present this mapping on a sample of 50,000 lines of extraction from two languages, English and Spanish (8,577 lines successfully uploaded and mapped at the time of submission). The lines are sampled from the data scoring above 0.7 metaphoricity. In the Drupal database, we can also store metaphor, frame, and construction information, which would otherwise be stored and viewable separately from the extraction data. This is clear in Figure 2 because the entries for source and target frames are live links (in blue). Because each extraction run contains potentially hundreds of thousands of lines of data, some of which is false positives, we sample across several extraction run outputs for metaphoricity values of over 0.7. This ensures that the data that makes it into the
Drupal database is more sure to contain true positives, without overburdening the data upload process.

4. Conclusion

A method for making metaphor extraction data from the MetaNet metaphor identification system publicly viewable is proposed, in the form of extraction run sampling and upload to a custom Drupal database. This method allows data to be easily searchable, and further enriched with more annotation, as desired by users. Query results can also be downloaded, and used for other purposes. The resulting database is a corpus of sentences (taken from myriad existing popular corpora, such as BNC) annotated for metaphors that can enable research into metaphor patterns across languages, across corpora, and across time.

References


Reliable measures of syntactic and lexical complexity: The case of Iris Murdoch
Stefan Evert, Sebastian Wankerl and Elmar Nöth
(Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany)

1 Introduction

Quantitative measures of the syntactic and lexical complexity of natural language text – such as type-token ratio (TTR), Yule’s K (1944) or Yngve depth (Yngve, 1960) – play a central role in stylometric analysis. They have been used to investigate stylometric differences between writers and settle questions of disputed authorship (Stamatatos, 2009), to explore the characteristics of translated texts (Volansky, Ordan, & Wintner, 2015), to identify determinants of style in scientific writing (Bergsma, Post, & Yarowsky, 2012), to study diachronic changes in grammar (Bentz, Kiela, Hill, & Butterly, 2014), to assess the readability and difficulty level of a text (Graesser, McNamara, Louworse, & Cai, 2004; Collins-Thompson, 2014), and as a feature in the multivariate analysis of linguistic variation (Biber, 1988; Diwersy, Evert, & Neumann, 2014).

In particular, several recent studies (Garrard, Maloney, Hodges, & Patterson, 2005; Pakhomov, Chacon, Wicklund, & Gundel, 2011; Le, Lancashire, Hirst, & Jokel, 2011) attempt to detect early symptoms of dementia in the last novels written by the British author Iris Murdoch, who was diagnosed with Alzheimer’s disease in 1997. These studies focus primarily on quantitative complexity measures, based on the assumption that beginning dementia reduces either the lexical or the syntactic complexity of a patient’s writing. Results were inconclusive: while the first two studies observed a promising decline of complexity in Murdoch’s last novel Jackson’s Dilemma published in 1995,1 Le et al. (2011) analyzed a larger sample of Murdoch’s writings and found that most of the quantitative measures did not to show any clear effects. In particular, they rejected the hypothesis of a decline in syntactic complexity.

Like most work in stylometry, all three studies fail to take the sampling distributions of complexity measures into account. As a result, they are prone to over-interpreting observed differences that may well be explained by random variation. Only Le et al. (2011) apply significance tests, but they test for a linear trend in complexity across the span of Murdoch’s writing career, which would not be consistent with the typical development of Alzheimer’s disease.

In this paper, we propose a novel methodology for the computation of reliable confidence intervals and significance tests for measures of linguistic complexity, inspired by ideas from bootstrapping and cross-validation. As an illustration, we apply the new method to the case of Iris Murdoch, showing that most of the differences observed in previous work are not significant and can indeed be accounted for by sampling variation.

2 Case study: The writings of Iris Murdoch

Iris Murdoch was one of the most renowned British writers of the post-war era. Between 1954 and 1995 she published a total of 26 novels and several non-fiction works on philosophy. Most of her novels were well received by literary critics and are known for their sophisticated topics (Spear, 1995). However, her last novel was received “without enthusiasm” and Murdoch

1And, for some measures, also in the penultimate novel published in 1993.
revealed that she experienced difficulties while composing the work (Garrard et al., 2005). Since this novel was published only a few years before her diagnosis, it is plausible to assume that her writing was already affected by the beginning dementia.

In our case study, we attempt replicate the results of prior studies by looking at 19 of Murdoch’s 26 novels, including the nine final ones. This provides complete coverage of Murdoch’s late work, which is of particular interest for the diagnosis of dementia and spans a period of almost 20 years. Our two main research questions are: (i) To what extent are the conclusions of prior research affected or even invalidated if sampling variation is taken into account? (ii) Does the onset of Alzheimer’s disease manifest in a clearly visible and significant decline of complexity (according to one or more of the quantitative measures)?

We purchased all 19 novels as e-books in epub file format, enabling us to extract the text directly without errors introduced by OCR software. We then used Stanford CoreNLP (Manning et al., 2014) for tokenization, sentence splitting, lemmatization, part-of-speech tagging, and syntactic parsing of the texts. Following Garrard et al. (2005) and Pakhomov et al. (2011), who excluded direct speech from their experiments, we flagged all such sentences in the novels. Detection of direct speech passages was greatly simplified by the epub format. However, the formatting of one of the e-books did not distinguish between quotation marks and apostrophes, so it had to be excluded from some of our experiments.

We report our findings for a wide range of complexity measures on the novels. Most of these measures consider lexical and syntactic aspects of the writing. From the lexical domain, we evaluate the vocabulary size and type-token ratio, the proportions of different word classes, Yule’s $K$ (which gives the probability of sampling the same word twice in a row) and Honoré $H$ (which measures the number of hapax legomena). In addition, we assess the approximate age-of-acquisition of the words appearing in a novel, using the word lists provided by Kuperman, Stadthagen-Gonzalez, and Brysbaert (2012) with a particular focus on the proportion of words learnt during later childhood and adolescence (beyond the age of 9 years). From the syntactic domain we evaluate simple measures like the average number of words or clauses per sentence. This is complemented by Yngve and Frazier depth, two measures which assign a higher weight to heavily left-branching sentences that are assumed to demand more working memory.

As a novel approach to measuring complexity, we also evaluate the use of statistical language models based on n-gram probabilities (Goodman, 2001). Such language models determine how well the text in one part of a novel can be predicted from the other parts of the novel. The perplexity of the model, a measure frequently used in speech processing, gives a good indication of the lexical and syntactic diversity of a text. This approach also has the advantages of being language-independent and not relying on complex linguistic pre-processing (Wankerl, Nöth, & Evert, 2016).

### 3 Bootstrapping confidence intervals and significance tests

Traditional significance tests based on a binomial sampling distribution cannot easily be translated to complexity measures, for two reasons: (i) they make the highly unrealistic assumption that a text (such as one of Murdoch’s novels) is a random sample of individual words or sentences; and (ii) they only apply to measures based on frequency counts or other numeric averages (such as mean sentence length). On the other hand, treating each text as a single item is often impracticable: in our case, it would be difficult to find any significant effects in a sample of size $n = 19$, and it would be impossible to test for a significant deviation of a single text.

---

2 *A Severed Head* (1961)
Oursolutioncombinestheideaofbootstrapping(Efron,1979)withthecross-validationprocedureappliedinmachinelearningandcomputationallinguistics. Wepartitionthesentencesofeachnovelintoconsecutivebinsof10,000tokenseach(leadingtodifferentnumbersofbinsondependingtonextlength). Leftovertokensattheendofthenovelarediscarded. Inourcasestudy,wegot $n \geq 6$ bins for all 19 novels.

Thecomplexitymeasureofinterestisthenevaluatedseparatelyoneachbin,resultinginnvalues $y_1, \ldots, y_n$ for a given text. Wecomputethemean

$$\mu_y = \frac{y_1 + \ldots + y_n}{n}$$

as an overallmeasurefortheentiretext, as well as the standard deviation $\sigma_y$ across the $n$ bins. Fora measurebasedonfrequencycountsoraverages, $\mu_y$correspondstothethatawouldbecomputeddirectlyfortheentiretext. Forthetype-tokenratio, $\mu_y$correspondstothestandardizedtype-tokenratio(cf.Kubát&Milička,2013), allowingameaningfulcomparisonoftextsofdifferentlength. Thesameholdsforotherlength-dependentmeasuresoflexicalcomplexity.

Thesamplingdistributionoftheoverallvalue $\mu_y$ canbedeterminedfromthestandard deviationacrossbins. Accordingtobootstrappingtheory, anapproximate95%confidence intervalfor $\mu_y$ isgivenby

$$\mu_y \pm 1.96 \cdot \frac{\sigma_y}{\sqrt{n}}$$

Thescalingfactor $\sqrt{n}$correctsforthelargervariabilityofmeasurementsinbinsof10,000 tokenscomparedtoentiretexts. Significance tests for different hypotheses canbeconstructedinasmilaryway, e.g. to determine whether asingle text differs significantly from the remaining body of work of an author, or whether there is a significant differencebetween two groups of texts.

As an example, Figure 1 shows $\mu_y$ with 95% confidenceintervalsfortwquantitative complexity measures: the ratio between nouns and verbs (left panel) and Honoré $H$ (right panel). Without confidence intervals, both plots would give a similar impression of a slight decline in complexitytowards the end of Murdoch’s writing career. However, the statistical uncertainty inherent in thenoun/verb ratio ismuch larger andits decline for the last two novels is not significant (as a rule of thumb, twotexts are significantly different if their confidence intervals do not overlap). Honoré $H$, bycontrast, is significantly lower in Jackson’s
Dilemma than in most other novels written by the author, suggesting that it might indeed reflect a cognitive impairment.

The remaining measures yield mixed results. None of the syntactic measures shows any significant change at the end of the writing career, in accordance with the conclusions of Le et al. (2011). From the lexical domain, the proportion of words acquired beyond the age of 9 shows a significant decline in the last novel (similar to Honoré H). The proportions of word classes do not show any significant trend. In particular, the number of pronouns fluctuates during the entire writing career, while the number of nouns remains more stable and increases only slightly towards the end. The number of adjectives increases in the beginning, remains stable for some time and slightly decreases at the end. The type-token ratio shows a non-significant decline in the last novel. The perplexity of n-gram models remains relatively low during the first third of the author’s writing career, then soars in the early seventies, and declines at the end. However, the confidence interval for the last novel is particularly wide so that no significant change can be observed.

References


Northern Haida (ISO 639: hdn) is a severely endangered language isolate spoken by fewer than 10 speakers in British Columbia and Southeast Alaska. The largest collection of original narrative texts in the language was recorded by John R. Swanton and published in 1908 using a pre-phonemic orthography. As part of a program of language documentation and revitalization in the community, we have been adapting Swanton’s original texts into the modern orthography (Hubert et al. 2016) and creating a linguistically-annotated digital corpus of approximately 100,000 words.

While much of Northern Haida grammar and lexicon (Enrico 2003, 2005) is well-documented in comparison to other Indigenous languages, no corpus studies have ever been undertaken on the language. In our project, we are exploring a range of basic questions about Northern Haida grammar to develop a fuller understanding of all aspects of the language.

While word order in Northern Haida is relatively rigid, speakers do have ordering choices in certain constructions, particularly those involving nouns and possessive pronouns (náay díinaa vs. gyáagan náay ‘my house’), as well as the general ordering of postpositional phrases with respect to nominal arguments in the clause. We will report on initial results of analyses of these variable-order constructions based on their frequency and distribution within the corpus.

In addition to improving our grammatical description of the language, we are also using the corpus to conduct forensic dialectology, examining the sub-phonemic alternations captured by Swanton’s original transcription. Of particular interest are certain dialectal shibboleths, such as the form of the definite suffix (-aay in the Alaskan dialect vs. -ee in the British Columbian dialect), which show clear community differentiation at later points in the 20th century, but which showed significant interspeaker and intraspeaker variation within Swanton’s corpus, allowing us to capture change in progress. We will discuss several such variations, and report on what insights the corpus allows us to glean about this earlier stage of the language.

We will conclude our presentation by looking at the real-world application of the corpus in supporting language revitalization in the community, showing how information from our analysis is being incorporated into improved language reference and teaching materials for younger learners.

References


If an atom is a letter, then a molecule is a word: applying corpus linguistic methods to chemistry

Maciej Eder, Michał Woźniak, Urszula Modrzyk, Rafał L. Górski (Institute of Polish Language, Poland) and Bartosz Grzybowski (Ulsan National Institute of Science and Technology, Ulsan, South Korea)

The buzzword quoted in the title of this paper, even if popular in chemistry, sounds very naïve for anyone who has some expertise in linguistics. Nonetheless, despite a rather shallow similarity between linguistics and organic chemistry on theoretical level, one cannot deny the usefulness of (some of) corpus linguistics methods to analyse internal molecule structures. This potential applicability makes such investigations feasible and desirable. The present paper demonstrates a procedure of identifying word forms – originally developed to assess Chinese script with no clear word boundaries – to divide complex chemical molecules into “meaningful” substructures. In this context, “meaningful” means groups of atoms that are local centres of reactions.

Let us assume that a molecule is a sentence (with some obvious caveats in mind, non-linearity of molecules being the most important one). If so, then a list of known molecules can be treated as a corpus. Quite striking is the fact that a commonly used convention of describing chemical structures (referred to as SMILES) uses sequences of characters, what makes any comparisons to corpora even more feasible. E.g., caffeine is coded as follows: CN1C=NC2=C1C(=O)N(C(=O)N2C)C.

Being one of the most crucial issues in organic chemistry, the question why certain groups atoms tend to keep together, while repelling others, has been approached using different methods, which are aimed at finding repetitive fragments of molecules. It can be assumed that methods derived from text mining can be adopted to (partially) solve the task.

If we name the “meaningful” groups of atoms as “words”, we need a device for finding them in our corpus since there are no explicit word boundaries. Grzybowski (2015) compared (in pairs) thousands of molecules, in order to extract their maximum common substructures, with the belief that they represent chemical “words”; this step was followed by a term frequency–inverse document frequency (tf-idf) heuristic. The elements which are obtained by this procedure behave as words in respect to Zipf’s and Heaps’ laws. Still, the linguistic motivation of this approach is rather weak, moreover it is computationally complex. Therefore we adopt a method for establishing word boundaries in Chinese, as proposed by Maosong et al. (1998).

The method uses the concept of sliding window, which divides a string of characters into chunks, and then computes an association measures inside the window as it moves. The association measure is a combination of two classical indexes: Mutual Information and t-tests.

Our preliminary tests, performed on a rather small corpus of 100,000 chemical
molecules, show that one can identify a considerably stable set of repetitive molecule parts that have well-defined “word-boundaries”. Certainly, the obtained results cannot be validated directly (there is no a priori definition of a chemical word). However, one can verify the results indirectly, e.g. via linguistic characteristics revealed by our “words”. In Fig. 1, an example of such a measure is shown, i.e. the classical frequency/rank dependence, followed flawlessly by our “corpus”.

References


tücan: a new web tool for sophisticated corpus linguistics
Daniel McDonald (Eberhard Karls University, Tübingen, Germany)

Analysis of large, digital collections of text is an increasingly prominent task within both academia and industry. Two major reasons for this are the increasing availability of born-digital data, and improvements in tools and methods for annotating and extracting useful information from natural language. Many of the cutting-edge developments for annotation and extraction take place within computational linguistics and natural language processing. Most of those working within these fields are familiar with the command line and computer programming. The kinds of tools and methods created and used within these fields, however, have a great deal of potential value elsewhere in linguistics, in lexicography, and in the social sciences and humanities more generally. In many of these traditions, practitioners have little experience with computer programming, and as such, are unable to put state-of-the-art developments to effective use. Software is therefore needed that can bring the power of computational methods into the hands of non-computational practitioners. The development of graphical and web-based interfaces are two possible strategies for bridging the divide. Often, however, such interfaces place hard constraints on what a researcher can accomplish: users have limited ability to iterate over structures within a dataset, or to work recursively. De Marneffe and Manning describe the importance of diversifying the user-base of currently available tools:

A major problem for the natural language processing (NLP) community is how to make the very impressive and practical technology which has been developed over the last two decades approachable to and usable by everyone who has text understanding needs. ... That is, usable not only by computational linguists, but also by the computer science community more generally and by all sorts of information professionals including biologists, medical researchers, political scientists, law firms, business and market analysts, etc. ...[T]he availability of high quality, easy-to-use (and preferably free) tools is essential for driving broader use of NLP tools (2008, p. 1).

Corpus linguistics is one of the main traditions that blurs the divide between computational and non-computational linguistics. Anthony et al. (2013) have described the important features in next-generation corpus linguistic software:

[Linguistic research] will rely increasingly on large corpora, advanced functionality, and sophisticated statistical methods. ...corpus tool development should be an open source initiative with tool components being developed in a modular fashion. By dividing tool components in this way, it becomes easier for tool functions and features to be extended, modified, or simplified depending on the need (pp. 155{156).
In this presentation, I provide an overview and demonstration of tücan (Tübingen Corpus Analysis suite), a new web application for investigating digital text collections, designed to address limitations in current corpus linguistic tools. Using corpkit (http://corpkit.readthedocs.io) as a backend, tücan allows users to upload and parse metadata-rich corpus texts, and then perform searches and visualisations of the parsed data structures and corpus statistics. Metadata can be encoded before or after annotation/parsing. A corpus of casual conversations, for example, can be annotated with speaker names (or other arbitrary metadata), using either of the conventions below:

ALYSSA: I suppose I could have one more
I didn't think you liked them. <metadata speaker="max">

Metadata can be used as search filters, or as symbolic subcorpora, so that users can view any search result within a three dimensional array, tracking lexicogrammatical feature frequency by subcorpus. Corpora can be automatically annotated based on query results, and annotations used to restrict or structure further searching. For instance, users can tag sentences matching some linguistic criteria, and include/exclude these sentences in future searches.

A second innovation of the tool is its ability to search for not only token level, but constituency and dependency level information as well. TGrep2 syntax can be used for searching of constituency trees, and a purpose-built syntax is available for dependencies. CQL queries are also recognised. Because the searching process is divorced from the process of displaying results, users can quickly generate novel views of the same data, switching between n-grams, collocations, lemma forms, POS tags or combinations thereof without re-interrogating the dataset.

Overall, the tool prioritises fluid movement through views into corpus data. Within the concordance interface, for example, users can sort or modify individual concordance lines. These edits automatically update other views, such as the frequency table and visualisations. New searches can be performed by selecting tokens of interest within concordance lines. In this way, workflows can be recursive or cyclical. The tool is free, open-source, and under version control. Login is possible for any researchers with a university affiliation via CLARIN. Its backend relies on popular, modern Python libraries for manipulating data (pandas/NumPy) and doing statistical work (scipy). Parsed data is stored in a CONLL-U-like format, with allows for arbitrary metadata representation. Parsing and interrogating operations can be parallel processed. The frontend relies heavily on Twitter Bootstrap for interface and d3js for plotting. In addition to the point-and-click interface, users are given a space into which they can enter natural-language like commands:

    search corpus for lemma matching processes: verbal
    show result with subcorpora as year and columns as lemma, pos
    sort result by increasing frequency
    plot result as line chart with title as "Example"

This expressive syntax allows the user to keep a precise log of previous steps taken during the process of data analysis.
A growing area of interest is the study of health-related discourse in online discussion forums. Such studies can be divided into those whose concern is risk talk online (e.g. Richardson, 2001; 2003) and those interested in online patient peer-support groups (e.g. Hunt and Harvey, 2015; Seale, 2006; Seale et al., 2010). This paper aims to contribute to the body of knowledge in these related fields. The paper presents a corpus-driven study of the uses of the expression (in the) real world in the five million word JABSforum corpus, a corpus comprising 3,272 forum threads gathered between 2005 and 2010 from the discussion forum of the JABS (Justice, Awareness and Basic Support) group website. The JABSforum corpus comprises what can be termed ‘vaccine-critical’ discourse, ‘vaccine-critical’ being the term used to describe organised groups which campaign against vaccination in general or for reform to government vaccination policy (Hobson-West, 2005; 2007). The JABS forum shares some features of the risk-talk forums described by Richardson (2001; 2003) and some aspects of patient peer-support forums, in that it serves as a place where parents can seek advice but is also a locus for organised resistance to vaccination policy.

The decision to look at the string (in the) real world arose from the observation that, with 65 citations, in the real world is among the more frequently occurring, linguistically meaningful, four-word N-grams in the JABSforum corpus. The adjective real is interesting. Tognini-Bonelli (1993:194-195) observes that real can serve a ‘focussing’ or a ‘selective’ function. When fulfilling a focussing function, real merely intensifies the noun (as in real difficulties). But when fulfilling a selective function, it is evaluative and encodes ‘inferential meaning’, that is, the choice of real + noun triggers the presupposition of the existence of another element which makes up a contrasting set and encodes an implicit negative evaluation of the latter (ibid.). The use of the adjective real is particularly interesting in the context of vaccine-critical discourse, in which appeals to the authority of the lived experience of parents are often used as a means for resisting mainstream medical-scientific opinion (Hobson-West, 2005).

The study reported here uses WordSmith 7.0 (Scott, 2016) to analyse JABSforum concordance data, although the starting point for the study was perusal of a list of 4-word N-grams from the JABSforum corpus calculated using AntConc 3.4.4 (Anthony, 2016). In order to establish the norms of use of (in the) real world in internet-based discourse, ukWaC is consulted, accessed by means of Sketch Engine (Kilgarriff et al., 2014). The relative frequencies per million words (pmw) of real world and in the real world in ukWaC and JABSforum are compared. A word sketch of real world is drawn up using Sketch Engine’s word sketch function in order to describe the typical uses of (in the) real world. LogDice is used as the measure of statistical significance. Finally, concordance evidence for (in the) real world in the JABSforum corpus is examined, using WordSmith Tools and the cluster program is used to identify frequent expressions. Frequency of recurrent strings in JABSforum is
expressed using raw frequency alone since WordSmith’s cluster program does not calculate statistical significance. In the discussion which follows, frequency information and collocational significance scores (where available) are shown in angle brackets. Examples of word forms are shown in lower case, lemmas in upper case.

A comparison of the relative frequencies of real world and in the real world in the JABSforum and ukWaC corpora reveals markedly frequent use of these expressions in the JABSforum corpus. The relative frequency of real world in JABSforum <19.75 pmw> is noticeably higher than in ukWaC <8.10 pmw> and that of in the real world even more so <13.65 pmw in JABSforum compared with 2.60 in ukWaC>. Concordance evidence from ukWaC shows that, in the expression real world, real fulfils a selective function. The implied contrast is with an imagined world. However, the nature of the evaluation implied varies according to the context in which real world is used. A Sketch Engine word sketch of real world in ukWaC shows that the expression typically occurs as part of a prepositional phrase. The most frequently occurring strings are LIVE in the real world <256; 3.72>, HAPPEN in the real world <91; 3.40>, WORK in the real world <88; 0.72>, EXIST in the real world <80; 3.34>, experience of the real world <35; 0.08>, out of touch with the real world <34; 3.07>, RETURN to the real world <30; 1.51>; OBJECT in the real world <28; 1.57>, OPERATE in the real world <27; 1.44> and APPLY in the real world <27; 0.97>. Expressions such as WORK/OPERATE/APPLY in the real world and object(s) in the real world are typically used in contexts in which the implied contrasting element is the world of (social) scientific ideas or theory. However, the world of theory is not usually evaluated negatively, except insofar as an idea is deemed worthy only if it is successfully applied in a real world context, as the following examples illustrate:

(1) ... hopes of producing an eco-friendly car that works in the real world.
(2) Judges looked for creative and unique business ideas that could be applied in the real world.
(3) Mandelbrot describes the use of fractal mathematics to understand objects in the real world.

In contrast, expressions such as LIVE in the real world, out of touch with the real world, and so on, are typically used to express a negative evaluation of a person or group of people by implying that their ideas or beliefs about the world are unrealistic. For example:

(4) They want their Government to live in the real world not a fantasy world.
(5) This is further evidence of how totally out of touch with the real world our politicians are.
(6) Many of us feel this plan is the product of ivory tower bureaucrats and is 1000 miles removed from the real world.

In the JABSforum corpus, the expression real world most frequently occurs in the strings work(s) in the real world <13 occurrences>, proven in the real world <6>, go/get (out) into the real world <6> and live/living in the real world <3>. Expressions involving (in the) real world are frequently used to express a negative
evaluation of the safety or efficacy of particular vaccines or of the robustness of the science behind them. Such claims are used by forum interactants to challenge the arguments of their interlocutors or of the medical-scientific community, as the following examples illustrate:

(7) ... show me the real science that Cervarix actually works in the real world rather than just believing the corporate marketing hype that sells this vaccine.

(8) At least green tea extract has been proven in the real world to both prevent and treat cervical cancer unlike Cervarix.

(9) Where is the statistical analysis of safety and side effects in the real world for combined dosing of vaccines?

Elsewhere, expressions such as go/get out into the real world and live/living in the real world, and so on, are used in order to express a negative evaluation of medical-scientific practitioners or other authorities:

(10) Why don't they get out into the real world and see what the real childhood health problems are?

(11) They dont live in the real world but in world of PR and PC that is all networked to common purpose.

(12) Anyone reading this article should try navigating the depths of the Department of Health's website [...] Out here in the real world of real parents, real teachers and real children we now all know at least one family with an autistic child.

In Example 12, above, it is interesting to note the way in which the lived experience of 'real' parents, teachers and children is evaluated positively in contrast with the claims of the Department of Health.

In conclusion, JABS forum interactants use appeals to the 'real world' in order to challenge arguments they encounter or to express a negative evaluation of the medical-scientific authorities. What is particularly interesting about the relatively high frequency in the JABSforum data of appeals to the 'real world' is that it reflects previous observations that appeals to the lived experience of parents is privileged in vaccine-critical discourse. Use of (in the) real world constitutes one of a number of rhetorical strategies used to resist vaccination policy.

References


Being clear – the politician’s riposte?
Ben Clarke (University of Portsmouth, UK)

In this paper, I explore the use of the adjective lemma *clear* in political and media-political discourse in the United Kingdom, flitting between corpus-based and text-analytic approaches in my analytical procedure. Taking first a corpus approach, I show that *clear*’s use in such political and media-political contexts increases diachronically over time towards the present day – particularly over recent time. This can be observed, for example, by studying the lemma’s attested use in The Hansard Corpus (Alexander & Davies, 2015), a diachronic corpus of UK parliamentary debates:

<table>
<thead>
<tr>
<th>Decade sub-corpus</th>
<th>Proportional frequency (instances per million words)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1800s</td>
<td>105.52</td>
</tr>
<tr>
<td>1810s</td>
<td>101.76</td>
</tr>
<tr>
<td>1820s</td>
<td>138.71</td>
</tr>
<tr>
<td>1830s</td>
<td>114.33</td>
</tr>
<tr>
<td>1840s</td>
<td>127.08</td>
</tr>
<tr>
<td>1850s</td>
<td>141.00</td>
</tr>
<tr>
<td>1860s</td>
<td>148.54</td>
</tr>
<tr>
<td>1870s</td>
<td>143.07</td>
</tr>
<tr>
<td>1880s</td>
<td>150.08</td>
</tr>
<tr>
<td>1890s</td>
<td>157.00</td>
</tr>
<tr>
<td>1900s</td>
<td>188.34</td>
</tr>
<tr>
<td>1910s</td>
<td>221.95</td>
</tr>
<tr>
<td>1920s</td>
<td>234.96</td>
</tr>
<tr>
<td>1930s</td>
<td>278.43</td>
</tr>
<tr>
<td>1940s</td>
<td>336.47</td>
</tr>
<tr>
<td>1950s</td>
<td>367.29</td>
</tr>
<tr>
<td>1960s</td>
<td>421.85</td>
</tr>
<tr>
<td>1970s</td>
<td>451.08</td>
</tr>
<tr>
<td>1980s</td>
<td>495.54</td>
</tr>
<tr>
<td>1990s</td>
<td>584.49</td>
</tr>
<tr>
<td>2000s</td>
<td>661.20</td>
</tr>
</tbody>
</table>

**Table 1: Frequency of lemma adjective ‘clear’ in The Hansard corpus**

Evidence to support the claim that this appears to be a trend specific to political discourse is offered by the fact that the lemma’s increase across time is only very moderate when studied in The Corpus of Historical American English (Davies, 2012), a corpus comprised of a number of different text genres concerning various subject-matter:

<table>
<thead>
<tr>
<th>Decade sub-corpus</th>
<th>Proportional frequency (instances per million words)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Moving to a text-analytic approach to clear’s use in present-day UK political contexts, I offer an analysis of BBC Radio 4 Today programme’s interview (14-06-2016) of the then Employment Minister, Priti Patel (IE below) by host Mishal Husain (IR below), shortly before the UK’s EU Referendum. I identify a seemingly emergent sense of clear as frequently used by Patel in this interview. The following are illustrative examples:

1. IR: how can you say as you do in this letter that (.). hh that you would ensure that universities scientists er farmers regional funds cultural organisations and others would continue to get the money that they get from the EU?

   IE: .hh well we’ve been clear that we would– we know that there’d be more than enough money to ensure that those .hh who currently get .hh that funding from the European Union would– that money would exist and it could still go there to those priorities

2. IR: the question I asked you was not about the contribution was– it was on what authority .h you as the Leave campaign make this promise it is (.). governments who make decisions on spending
IE: well we are saying that obviously .h government can choose after the referendum in terms of how it would .hh spend the money if we take back control of our money from the European Union .hh and we have been abundantly clear that there would be more than enough money .hh to ensure that those who now get funding from the EU including universities scientists farmers .h regional (funds) .h erm would continue to [get ] money

This use of clear is characterised by the combination of a number of particular lexicogrammatical and wider discursive features: a first person grammatical Subject; a past perfect main verb phrase; ‘that’ relativiser in R1 position which therefore introduces an embedded clause as an adjective phrase post-modifier; reference in the immediately preceding discourse (often of the IR) to the IE’s past communicative acts – or those of the party or the collective that the IE represents (e.g. say, this letter, the question I asked you, this promise); etc. Especially, clear in this relational: attributive transitivity construction, in Halliday’s (1994) terms, seems marked where used with an animate grammatical Subject and post-modified with a ‘that’-relative clause.

The aforementioned findings from this close, textual analysis are used to inform subsequent diachronic corpus inquiries using The News on the Web Corpus (NoW), a monitor corpus of online news starting in 2010. This second-step corpus analysis reveals that the particular aforementioned use of clear is indeed a very recent language phenomenon (see Arts et al., 2013 on short-term diachrony; cf. what is revealed by the earlier Hansard Corpus analysis), all but exclusively used in the NoW corpus in news texts with a political subject-matter.

In bringing the paper to a close, I speculate that the general increase in the use of clear – and particularly the specific sense focused on in my paper – functions for politicians as a push-back against claims of their evasiveness in response to direct questioning in public-facing discourse (Harris, 1991; Bull, 2008) as well as against other events which have undermined the public trust in politicians (e.g. the MPs’ expenses scandal of 2009: The Telegraph, 2009). In order to do so, I consider the wide-span collocational behaviour of I/we + have been + clear that + CLAUSE sense of clear when in use, which attests collocates indexing past communicative activity.

References


Making sense of multivariate analyses of linguistic variation
Stefan Evert (Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany)

Multivariate analysis, which exploits correlational patterns among a large number of linguistic variables, has been established as an important and successful technique in many subfields of corpus linguistics such as register variation (Biber, 1988), dialectology (Speelman, Grondelaers, & Geeraerts, 2003), translation studies (De Sutter, Delaere, & Plevoets, 2012) and authorship attribution (Jannidis, Pielström, Schöch, & Vitt, 2015). In this approach, texts (or other linguistic samples) are represented as high-dimensional vectors of quantitative features, the distance between vectors is understood as an indicator of linguistic dissimilarity, and a small number of latent dimensions are identified to capture the main patterns of variation in the data (usually corresponding to correlations between large groups of features). Individual studies differ in their choice of quantitative features – ranging from measurements grounded in a specific linguistic theory (Diwersy, Evert, & Neumann, 2014) to plain frequency counts in a “bag of words” approach (Jannidis et al., 2015) – and in the particular mathematical algorithm used to identify latent dimensions – usually an unsupervised technique such as principal component analysis (PCA; Diwersy et al., 2014), correspondence analysis (CA; De Sutter et al., 2012) or factor analysis (FA; Biber, 1988); some authors also apply linear discriminant analysis (LDA; Baayen, van Halteren, Neijt, & Tweedie, 2002) or another supervised algorithm to exploit additional information about the texts.

A methodological key problem lies in the difficulty of making linguistic sense of the results of a multivariate analysis. Typical approaches include (i) a hermeneutic interpretation of the weighted feature combinations corresponding to each latent dimension, based on human intuition, (ii) visualizing the average coordinates of external categories (such as text types, authors or translated vs. original texts) in the latent dimensions, or (iii) comparing an unsupervised clustering of the text vectors to these external categories. Such attempts are prone to over-interpretation and researcher bias (i), fail to show whether the features contributing to a latent dimension are correlated or complementary (i), or they establish that a multivariate analysis differentiates successfully between external categories, but do not explain how these differences arise from the original features (ii, iii).

This poster explores novel approaches to the linguistic interpretation of multivariate models. First, the feature weights of a latent dimension should not be taken at face value: researchers need to consider their statistical uncertainty (determined by cross-validation or bootstrapping) as well as the distribution of feature values (especially wrt. external categories such as text types). Second, relevant features can be identified by measuring their contribution to the separation of unsupervised clusters or to groupings based on external categories (using techniques such as random forests or recursive feature elimination). Third, a secondary multivariate analysis within clusters or groups reveals how different features combine in correlated or complementary ways into a latent dimension, giving deeper insights about the interactions between individual features.

The new approaches are illustrated with examples from authorship attribution and translation studies. R code implementing the case studies will be made available at http://www.stefan-evert.de/PUB/Evert2017CL/.
References


Mining a corpus of online hotel reviews: a pilot study
Phoebe Lin and Amy Suen (The Hong Kong Polytechnic University, Hong Kong)

Electronic word-of-mouth (eWOM) is vital in today’s competitive hospitality market. It is even more important to luxury hotels, which set hospitality standards worldwide. Online reviews on TripAdvisor, for example, are influential because they have a large readership. Studies have also shown their immediate impact on booking intentions (Vermeulen & Seegers, 2009; Ye et al., 2011).

This paper reports the findings of a corpus-driven study which examined 20,395 reviews about 12 Hong Kong 5-star hotels on TripAdvisor. The study aims to address two research questions about how the 5-star hotels reply to reviews while striving to promote their luxury brand image. These two research questions are: 1) how often do hotels reply to reviews? and 2) how do hotels reply to positive and negative reviews? Data analysis consists of two stages: quantitative analyses of reply counts and computer-assisted analyses of the language of the replies.

The results indicate that hotels reply more often to negative than positive reviews (reply rates of 65% vs 45%). There has also been a radical change in ‘reply culture’ over the past 12 years (2004-2016). In 2008, no hotel replied to reviews. However, in 2016, hotels replied to almost every review posted on TripAdvisor.

These replies, totalling 56,000 words, were then compiled into two corpora (replies to positive vs negative reviews). Keyness analyses were then applied (using Wmatrix) to reveal lexical, grammatical and semantic features that distinguish replies to negative reviews from replies to positive reviews. Amongst other findings, we noted that replies to low rating reviews are characterised by a tone of regret, ownership and reassurance. However, replies to high rating reviews are characterised by an affective and informal tone, as well as reiterations of the hotels’ competitive advantages.

This study provides a corpus-driven description of the ways in which 5-star hotels respond to online reviews. It highlights the various linguistic resources that are often exploited in the online review environment. These findings may shed light on future research on communication strategies in tourism and hospitality management, and inform the development of relevant ESP teaching materials.

References

Formulaic sequences are an umbrella term for conventionalised sequences of two or more words that form holistic functional, meaning and/or processing units. Common types of formulaic sequences, such as idioms, proverbs, speech formulas, and collocations, often display qualities of formal fixedness, semantic non-compositionality, and high frequency of occurrence to different extents. Over the last decade, however, there is a growing interest in the prosodic fixedness of formulaic sequences, including their tendency to align with pauses (Lin & Adolphs, 2009; Erman, 2007; Wray, 2004), their fast speech rate compared with non-conventionalised word sequences (Lin, 2010) and their restricted tonal patterns (Ashby, 2006; Lin, 2013).

The lack of spoken corpora with prosodic annotation or alignment with audio/video streams, however, has hindered the growth of research on the prosodic patterns of formulaic sequences (Lin & Chen, forthcoming). This paper presents a new computer tool developed recently to tackle the shortage of multimodal data and offer researchers a new tool for exploiting the Web as multimodal corpus. Using the online interface, users may compile and concordance their own large, sustainable and dynamic YouTube corpora. While the present paper demonstrates the use of the tool for profiling the prosodic patterns of formulaic sequences, the tool can support any research into the interfaces between lexis, prosody and gestures in naturally occurring spoken discourse.

References

**Lexical bundles in academic bio-data: A corpus interdisciplinary analysis**

Kathy Ling Lin (Shanghai Jiao Tong University, China) and Isaac Nuokyaa-Ire Mwinlaaru (Hong Kong Polytechnic University, Hong Kong)

In the investigation of academic English, compared with the much-studied genres such as research articles and degree theses, academic bio-data is an underexplored “para-genre” or “para-text” (Genette, 1987[1991]). This “para-genre”, however, is very important and closely related to an academic’s identity construction (i.e., claim membership of a disciplinary discourse community while presenting his or her own professional identity). To bridge the gap and establish its structural identity (or in Bhatia’s (1993) term “generic integrity”), Mwinlaaru (in press) innovatively gave an SFL-based account of its generic structure by analyzing 200 biodata written by applied linguistics scholars, with 100 each from journal articles and seminar posters. The present project is a further extension in this line to examine cross-disciplinary similarities and differences in lexical bundle use in academic bio-data, and to associate the findings from bundles analysis with their structural variations and peculiarities across disciplinary boundaries.

To fulfill the research purpose, the authors collected 300 academic bios from 15 high-ranking prestigious journals in three disciplines based on Journal Citation Database and the recommendations from our disciplinary informants. The corpus of academic bio statements includes 100 each from Applied Linguistics (AL), Industrial Engineering (IE), and Physical Sciences (PS). The source journals for the three disciplines are: for PS, *Solid State Nuclear Magnetic Resonance* (SSNMR), *CHEMPHYCHEM* (CHEM), *The Journal of Physical Chemistry Letters* (JPCL), *Soft Matter* (SM), and *Chemical Physics Letters* (CPL); for AL, *Discourse & Society* (D&S), *Discourse Studies* (DS), *English for Specific Purposes* (ESP), *Language & Learning* (L&L), and *Text & Talk* (T&T); and for IE, *IEEE Transactions on Engineering* (IEEETE), *IIE Transactions* (IIEIT), *Industrial Management & Data Systems* (IM&DS), *Journal of Manufacturing Systems* (JMS), and *Journal of Product Innovation Management* (JPIM). The average lengths of academic bio statements for the three disciplines are respectively 73.9 words (Applied Linguistics), 88.7 words (Industrial Engineering), and 85.6 words (Physical Sciences). Before we conducted bundles analysis, rhetorical structural theory and SFL genre-based approach have been applied to describing the typical structure of this genre and their related cross-disciplinary variations, which could help account for the differences and the featured use of lexical bundles across the three contrasting disciplines. The findings from the structural analysis show that while Applied Linguistics scholars favour a two-tier contextual structure and an inventorying style of presentation in bios, Physical Science scholars prefer a three-tier structure and a chronicling mode. Industrial Engineering scholars lie between the two, preferring a three-tier structure and a synthesis of chronicling and inventorying styles. These disciplinary variations in the
rhetorical structure of biodata are realised by different choices in phraseological patterns in the bios.

In phraseology study, 4-word bundles analysis was conducted, as “the four-word scope is the most researched length for writing studies...manageable size for manual categorization and concordance checks” (Chen & Baker, 2010: 32). AntConc 3.4.4w was used, with cut-off points determined based on our data observation and a very close reading of the rich literature on bundles analysis (the cut-off standardized frequency: 0.6 times per thousand words, the raw cut-off frequency: 4 for AL, 5 for PS and IS; distribution: 5%). The study shows interesting cross-disciplinary commonalities and differences in four-word lexical bundle use in terms of their structure and function. In our talk, we will present in detail our findings on bundles analysis, e.g., the most frequently-used lexical bundles and their categories in terms of function and structure, to see how disciplinary variations in the rhetorical structure of biodata are realised by different choices in phraseological patterns in the bios. As an illustration, we made a table to show the findings from the comparative study of lexical bundle use in AL and PS bio-data (see Table 1). More interesting findings will be presented in our talk. Research contributions and implications for ESP teaching and research will be discussed.

Table 1 A case study: use AL and PS for a comparison

<table>
<thead>
<tr>
<th>Bundles (tokens, range) in AL bios (14 types, 132 tokens)</th>
<th>Bundles shared by both disciplines</th>
<th>Bundles (tokens, range) in PS bios (39 types, 342 tokens)</th>
</tr>
</thead>
<tbody>
<tr>
<td>at the university of (40, 39), in the department of (15, 15), her research interests include (13, 13), in the school of (9, 9), research interests include the (7, 7), his research interests include (6, 6), is the author of (6, 6), university of hong kong (6, 5), from the university of (5, 5), is a senior lecturer (5, 5), is an associate professor (5, 5), is associate professor of (5, 5), is senior lecturer in (5, 5), the school of English (5, 5)</td>
<td>at the university of, in the department of, from the university of</td>
<td>at the university of (38, 30), under the supervision of (22, 19), from the university of (21, 20), his ph.d in (18, 18), received his ph.d (15, 15), his research interests include (11, 11), in the department of (11, 8), the department of chemistry (11, 9), the university of california (10, 9), a ph.d in (9, 9), degree in chemistry from (9, 8), in chemistry from the (9, 9), ph.d in chemistry (8, 8), the max planck institute (8, 7), his ph.d from (7, 7), ph.d from the (7, 7), ph.d in # from (7, 7), received her ph.d (7, 7), received his b.s (7, 7), b.s degree in (6, 6), he is currently a (6, 6), he received his ph (6, 6), her ph.d in (6, 6)</td>
</tr>
</tbody>
</table>
References


The JSI Newsfeed corpus is a new family of web corpora created from the JSI newsfeed created by Jozef Stefan Institute, Slovenia (Trampus et al 2004). JSI newsfeed is a clean, continuous, real-time aggregated stream of semantically enriched news articles from RSS-enabled sites across the world. The newsfeed is available in many languages including Arabic (250 million words), Catalan (40 million), Czech (146 million), German (936 million), English (8 billion), Finnish (75 million), French (782 million), Croatian (150 million), Hungarian (77 million), Italian (335 million), Korean (130 million), Dutch (176 million), Polish (150 million), Russian (500 million), Spanish (1.5 billion), Serbian (38 million), Swedish (147 million). The project continuously processes 80,000 RSS feeds which bring between 350,000 and 600,000 articles every day. All articles are cleaned so that the main body of text is included, duplicated articles are removed and, most importantly, all data is time stamped.

The JSI newsfeed has been processed into the JSI Newsfeed corpus by Sketch Engine (Kilgarriff 2014). Sketch Engine is a corpus query software with automated corpus building tools. Sketch Engine currently holds 150 TB of data including 400 corpora in 85+ languages. Sketch Engine specializes in processing extremely large corpora with a size of up to dozens of billions of words.

The JSI Newsfeed corpus was tagged for parts of speech and the time stamps were used to augment the corpus with diachronic annotation. Currently, the corpus covers the time period from the year 2014 onwards. The feeds are crawled continuously by Jozef Stefan Institute and the corpus is amended daily with the latest articles. By combining this data with other English web feed corpora, a total period from 2009 up to the current day is covered.

The diachronic annotation is extremely valuable in connection with Sketch Engine and its trends feature. The trends feature analyses the frequency of the use of a word in time by comparing the frequency of use across a series of comparable time periods.

The algorithms will then present the user with a list of words which have changed its frequency of use over a longer time period, i.e. neologisms candidates and words going out of use.
The words listed on the results screen may contain certain items which do not qualify as words whose use changed over time, typically, misspelt words might fall into this category. This, however, does not reduce the usefulness of such a tool. Even with a certain level of pollution, it will take a lexicographer only fraction of time to discover neologisms compared to the traditional approach of reading media trying to discover a
new word. The results are sorted by the trend value, i.e. the words with the biggest change are at the top irrespective of whether the change was positive (=growing usage) or negative (=decreasing usage).

In Sketch Engine, the trend value can be calculated using two methods: either a simple linear regression is used, or the Mann-Kendall test. The latter is used because it showed to be more robust with noisy data, such as very imbalanced time slices, or random local extremes, both of which occur very often especially with web corpora. Methods of calculating trends are discussed in great detail in (Kilgarriff, 2015) and (Herman, 2013).

Apart from the trend value, the p-value is calculated and shown, and for each item in the list the interface allows the user to show concordance lines and frequency distribution according to the time periods.

The trends feature on its own cannot be used to identify changes in word senses. The current research focuses on exploiting the diachronic annotation even further in the direction of an automatic identification of the semantic shift. The approach combines word sketches, one-page summaries of a word’s collocational behaviour, with the trends feature. Preliminary research (Baisa, 2015) shows that attention should be paid to the change of the collocational behaviour of the word in time because changes in collocational behaviour correlate with word sense change. This could lead to a new functionality in Sketch Engine aimed at automatic identification of the semantic shift with the help of diachronically annotated corpora.

References

Herman & Kovář (2013). Methods for detection of word usage over time. RASLAN 2013 Recent Advances in Slavonic Natural Language Processing, 79.
Corpus-assisted discourse analysis of family legislation changes in Slovene media

Senja Pollak (Jozef Stefan Institute, Slovenia) and Ana Marija Sobočan (Faculty of Social Work, University of Ljubljana, Slovenia)

1 Introduction

Corpus linguistics and (critical) discourse analysis have been, with the aim of uncovering and questioning the representations of different social groups or phenomena, successfully combined in several studies (e.g. Krishnamurthy, 1996; Stubbs, 1998; Teubert, 2000; Baker, 2006; Mautner, 2009; Baker & Lavon, 2015). In our contribution we present a corpus-based discourse analysis of media coverage of the Slovene Act Amending the Marriage and Family Relations Act (hereby ‘the Act’). We analyse and compare the discourses in three online and printed newspapers, Demokracija, Dnevnik and Mladina that were chosen based on their diverse standpoints towards the topic of interest. The selection of terms for analysis is based on terminology extraction, which by a statistical measure for comparing the terms of a domain corpus to the terms of the reference corpus, identifies the single- or multi-word terms that are characteristic for a selected corpus. In the first study (see Section 4) we analyse the concordances of the most frequent multi-word terms of the entire corpus, while in the second study (Section 5) we analyse the terms that are distinctively characteristic for each medium.

2 Corpus description

The corpus consists of 69 texts from the three online and printed media. The three are generally recognized as ideologically oriented as right-wing / conservative (Demokracija), centre (Dnevnik) and left-wing / progressive (Mladina). The texts were selected based on the relevance to the topic and limited by the studied period spanning from the 15th of November 2015, a few weeks prior to the referendum on the Act and to the 31st of December 2015, a week after the referendum.

<table>
<thead>
<tr>
<th></th>
<th>Mladina</th>
<th>Dnevnik</th>
<th>Demokracija</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td># articles</td>
<td>10</td>
<td>27</td>
<td>32</td>
<td>69</td>
</tr>
<tr>
<td># sentences</td>
<td>493</td>
<td>876</td>
<td>1836</td>
<td>3205</td>
</tr>
<tr>
<td># word tokens</td>
<td>9,492</td>
<td>17,063</td>
<td>36,193</td>
<td>62,748</td>
</tr>
</tbody>
</table>

Table 1: Corpus statistics.

1 Term ZZZDR, the abbreviation for the Act, was used as a query in the archives of the three newspapers. The returned search results (articles were downloaded in October and November 2016) were manually filtered to confirm the relevance to the topic.)
In total 69 articles corresponded to the defined criteria (see corpus statistics in Table 1). The texts were pre-processed with the ToTrTaLe tool (Erjavec, 2011) including sentence and word tokenisation, lemmatisation and annotation with morpho-syntactic descriptions. The text was uploaded to the SketchEngine corpus linguistic tool and concordancer (Kilgarriff et al., 2014).

3 Media discourses on family legislation changes (Act Amending the Marriage and Family Relations Act)

The study focuses on media discourses related to the ‘Act amending the Marriage and Family Relations Act’, which aimed at changing the then current family legislation. The first attempt for changes happened in 2009, when the Ministry of work, family, social affairs and equal opportunities proposed the ‘Family Code’, which was supposed to substitute the Marriage and Family Relations Act originating in the 1970s (hence outdated). The Code was adopted by the Parliament, but rejected with a referendum in 2012. Already then, there was much public attention and media focus on the proposed legislative changes, and the interest of the public was mainly sparked by the legislation introducing pluralisation of families and equalising biological (parents defined by a genetic link to children) and social parenthood (parents defined by a social, care-based link to children), by equalising the rights and status of same-sex and different-sex partnerships and parental rights. In December 2014 the ‘Act amending the Marriage and Family Relations Act’ was filed in parliamentary procedure and was adopted, implementing a symbolic and legal equality for families of partners of same or different gender. As the Code, also the Act was rejected with a referendum (on 20th December 2015).

Other authors have already analysed the discourses related to family legislation in Slovenia, most notably Vezovnik (2015), who published an exhaustive analysis of argumentation schemes and structures related to the ‘Family Code’, while Kuhar (2015) published an analysis of populistic strategies of the opponents of the ‘Family Code’. To the best of our knowledge, we are the first to address the context of the last referendum and the ‘Act amending the Marriage and Family Relations Act’ and to analyse the discourses with combined quantitative and qualitative methods (i.e. as a corpus assisted discourse study). Studying particularly the discourses on (changes of) family legislation proves to be an interesting and relevant endeavour, as it can be observed that numerous societal norms and values are reflected, reproduced or refuted in this framework.

4 Discourse analysis of top ten terms

Using the online term extraction workflow (Pollak et al., 2012), based on a comparison of noun phrases in domain and reference corpus, we extracted a list of terms ranked by their termhood score. In our case, the domain corpus is the corpus of 69 articles from the three selected media, while the reference corpus is the 100-million word corpus Kres (Logar 2012), containing Slovene texts of various genres, from daily newspapers, magazines, books (fiction, non-fiction, textbooks), web pages, and similar, with a balanced genre structure. We analysed the top 10 extracted bigrams, presented in Table 2. Since the corpus is imbalanced, the largest subcorpus
(Demokracija) predominantly impacts the term extraction. For example the term theory of gender appears nearly exclusively (72 out of 75 occurrences) in Demokracija.

<table>
<thead>
<tr>
<th>Bigram rank by termhood value</th>
<th>Term translation</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>marriage</td>
<td>238</td>
</tr>
<tr>
<td>2</td>
<td>amending act</td>
<td>137</td>
</tr>
<tr>
<td>3</td>
<td>family relationship</td>
<td>96</td>
</tr>
<tr>
<td>4</td>
<td>theory of gender</td>
<td>75</td>
</tr>
<tr>
<td>5</td>
<td>family code</td>
<td>58</td>
</tr>
<tr>
<td>6</td>
<td>human rights</td>
<td>72</td>
</tr>
<tr>
<td>7</td>
<td>same-sex partnership</td>
<td>40</td>
</tr>
<tr>
<td>8</td>
<td>same-sex couple</td>
<td>38</td>
</tr>
<tr>
<td>9</td>
<td>act amending the marriage and family relations act</td>
<td>32</td>
</tr>
<tr>
<td>10</td>
<td>constitutional court</td>
<td>53</td>
</tr>
</tbody>
</table>

Table 2: 10 top-ranked bigram terms.

We analysed all corpus examples that contained the selected terms and interpreted them from the perspective of critical discourse analysis. We identified six discursive frames:

- family as ‘natural’ and heterosexual relationship as the norm
- theory of gender
- dissolution of ‘Family’ and values, a threat to the social order
- hegemony of the ruling elites, manipulations and state of mind
- human rights, women’s rights and children’s rights
- civil rights, nation building and nationalism

The analysis shows that some terms (e.g., ‘theory of gender’) and discourses significantly mark only the texts in one of the media (Demokracija) and that certain terms (e.g. human rights) are used in conflicting discourses (i.e. both proponents and opponents of marriage equality are portrayed as human rights violators). Significantly, the homophobic, racist, and nationalist discourses of opponents of family legislation changes are producing moral and sexual panics (moral panics specifically related to issues of sexuality and reproduction).
5 Discourse analysis of contrasting terms

As in the selection presented in Section 4, we used the term extraction workflow (Pollak et al., 2012), but here we performed the term extraction for each subcorpus (medium) separately. Since we were interested in contrasting terms, characteristic for each of the subcorpora, we compared the lists of 150 top-ranked terms of each subcorpus and kept for analysis only the terms (n-gram noun phrases) that were listed in only in one of the three term lists. We grouped the contrasting terms in different semantic fields:

- partnership relationships
- legislation and referendum
- social semantic fields: gender, children, upbringing, nationality, human rights, values
- stakeholders / agents
- meta-discourse

We compared the differences in framing the referendum of three media. Different terms of the three media are characteristic for each field. For example, the differentiating terms selected from Demokracija under stakeholders category contain family members (husband, father, grandparent, mother, grandchild, wife, mom) that are presented in heteronormative matrices of an essentialist conception of family.

6 Conclusions

The corpus-driven analysis of the discourses of the three media has revealed the discourses and differences in reporting in three online and printed media reporting on the family legislation changes and the referendum. Standing out are the discourses propagating homophobia, islamophobia, anti-feminism, etc. that were the dominating discourses especially in the conservative medium Demokracija, and contributed to creating an atmosphere of danger and fear. By using term extraction, we limited our analysis on characteristic nouns (or noun phrases with a noun as a head word), but in future work we will extend the analysis to other parts-of-speech.

References


This study investigates determiner usage by Chinese learners of English, taking account of the Fluctuation Hypothesis by Ionin, Ko, and Wexler (2004), which states that learners’ article choice fluctuates for mixed settings of the semantic parameters definiteness and specificity. In an attempt to combine corpus-linguistic and cognitive approaches, the study employs a dual methodology: corpus data is complemented by data collected through a forced choice elicitation test. The forced choice elicitation test data was obtained using an online survey with 39 participants, most of whom are students at Sun Yat-sen University. The corpus data, from which three 100 sentence random samples were analyzed, was taken from the Sun Yat-sen University Corpus (SYSU-C), compiled by Küchler (2015).

The semantic concept of definiteness can be found in most languages. However, the way this concept is grammaticalized in different languages varies strongly and the realization can take various lexical, morphological or positional forms. Some languages, such as English or German, have a fully functional grammatical category of articles that encode definiteness, while in contrast, other languages, such as Samoan, encode specificity. Other languages, such as Korean or Russian, do not have an article category. Definiteness in these languages is expressed by different means, for example demonstratives.

In the area of first language acquisition, there has been strong evidence suggesting the existence of unconscious knowledge of language that lead Chomsky to postulating the notion of Universal Grammar (Chomsky, 1965, 1986). In addition, there has been research that suggest an effect of UG and its parameter setting model in second language acquisition as well (see Schwartz and Sprouse, 1994, 1996). Investigations of this effect revealed that L2 learners of English whose L1 does not have a fully functional grammatical category of articles have difficulties acquiring the correct usage of the English article system. Ultimately, that lead to the postulation of the Fluctuation Hypothesis (Ionin, 2003; Ionin et al., 2004), which states that until learners have received sufficient input to set the article choice parameters correctly, they will fluctuate in certain environments of these parameters.

Based on the assumption of the learners’ full access to universal grammar, Ionin et al. propose the Article Choice Parameter, presented in (1), that determines how articles are selected in languages. While English (neglecting the colloquial use of referential this) has the first setting, there are other languages, like Samoan, that have the second setting (Ionin et al., 2004, p. 12).

(1) The Article Choice Parameter (for two-article languages) A language that has two articles distinguishes them as follows: The Definiteness Setting: Articles are distinguished on the basis of definiteness. The Specificity Setting: Articles are distinguished on the basis of specificity.

(Ionin et al., 2004, p. 12)
Resulting from the two possible settings of the Article Choice Parameter, Ionin et al. predict that learners go through a period of fluctuation before sufficient language input leads them to setting the parameter correctly. The fluctuation is likely only to occur for mixed parameter values ([+definite] [-specific] and [-definite] [+specific]) and not for either positive or negative parameter values ([+definite] [+specific] and [-definite] [-specific]). These predictions, as seen in table 1, are consistent with Lyons (1999), who argued that languages with articles mark either definiteness or specificity. According to the definition of the Fluctuation Hypothesis given below, the learners’ interlanguage grammars are constraint by UG, however a single learner might have two rivaling interlanguage grammars in the process of language acquisition.

(2) The Fluctuation Hypothesis (FH) for L2 article choice:
1. L2-learners have full access to the features that can underlie article choice cross-linguistically: the features [+definite] and [+specific].
2. L2-learners fluctuate between dividing English articles on the basis of definiteness vs. specificity, until the input leads them to choose the definiteness option.

(from Ionin et al. (2009), based on Ionin et al. (2004, p. 8))

As Ionin et al. note, the explanation that the FH offers for non-random article errors is, however, limited to contexts where no other features, such as partitivity, interact. This is mirrored in their means of data collection, as the forced choice elicitation tasks carefully controlled for these circumstances.

<table>
<thead>
<tr>
<th>[+definite] (target: the)</th>
<th>[-definite] (target: a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[+specific] correct use or the</td>
<td>overuse of the</td>
</tr>
<tr>
<td>[-specific] overuse of a</td>
<td>correct use of a</td>
</tr>
</tbody>
</table>

Table 1: Predictions for Article Choice in Chinese L2 English (from Ionin et al. (2004))

Related to article misuse, as predicted by the Fluctuation Hypothesis, article omission, which can also be described as overuse of the zero article, is frequently reported. Different studies in SLA have found that even advanced learners whose L1 lacks articles exhibit an overuse of the zero article in their interlanguage grammars (Bergeron-Matoba, 2007, p. 7). To account for this phenomenon, there are two opposing views. On the one hand, there are researchers, utilizing a no UG access approach, arguing that the variability in the interlanguage grammars is due to some form of either temporary or permanent impairment (Prévost and White, 2000, p. 108). On the other hand, there are researchers, following an approach permitting UG access, claiming that the interlanguage grammars contain abstract functional categories and features. The variability found in empirical studies results from difficulties mapping that abstract knowledge to the surface morphological manifestation (Prévost and White (2000, p. 108), following Haznedar and Schwartz (1997)). They state that:
the Missing Surface Inflection Hypothesis (MSIH) proposes that L2 learners have unconscious knowledge of the functional projections and features underlying tense and agreement. However, learners sometimes have a problem with realization of surface morphology, such that they resort to non-finite forms [...]. (Prévost and White, 2000, p. 103)

Combining Full Transfer/Full Access (Schwartz and Sprouse, 1996) with missing surface inflection (Haznedar, 2001), the MSIH states that the initial state of the L2 grammar is the final state of the L1 grammar, which then needs to be restructured as it fails to account for input data. In contrast to first language acquisition, the activation of functional categories is dependent on the acquisition of morphology (Haznedar, 2001, p. 279). Hence, issues in L2 production stem from morpholexical aspects rather than systematic syntactic deficits (Haznedar, 2001, p. 280). This hypothesis offers a possible explanation for article omission errors a posteriori and has therefore been debated controversially (Granfeldt, 2000).

The learner groups that have been researched extensively in previous studies include Korean (Crosthwaite, 2014; Ionin et al., 2009), Russian (Chrabaszcz and Jiang, 2014; Ionin, 2003; Ionin, Zubizarreta, et al., 2008) and, to a lesser extent, Turkish (Dikilitaş and Altay, 2011), Japanese and Chinese native speakers (Snape et al., 2006). However, especially L1 Chinese learners are of particular interest, since there is research suggesting that the grammatical category of articles is in the process of grammaticalization in modern Chinese (P. Chen, 2003, 2004; C.-T. J. Huang, 1987; Li and Bisang, 2012). Consequently, it would be interesting to see, how Chinese learners of English perform in their article choice in comparison to other learners with an article-less L1 as well as to learners whose L1 does have articles (c.f. Snape et al., 2006; White, 2008).

While most of these studies have used forced choice elicitation tasks as means of data collection, it might be interesting to study the usage in academic writing as well, because it is rather natural language that was not collected in a focused test. Furthermore, the overall frequency of articles is highest in academic English across all genres (Biber et al., 1999, p. 267). Moreover, depending on the corpus used in the analysis, it can be investigated in how far some variation in the usage patterns can possibly be attributed to different levels of proficiency.

<table>
<thead>
<tr>
<th></th>
<th>[+definite] (target: the)</th>
<th>[-definite] (target: a)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>L1 Chinese</strong> (n=37)</td>
<td>the</td>
<td>a</td>
</tr>
<tr>
<td>[+specific]</td>
<td>84 %</td>
<td>14 %</td>
</tr>
<tr>
<td>[-specific]</td>
<td>89 %</td>
<td>9 %</td>
</tr>
</tbody>
</table>

Table 2: Total Article Choice

The results of the analysis of this study do not indicate fluctuation in the learners’ article choice as predicted by the Fluctuation Hypothesis, with low overall error frequencies (c.f. Table 2). In more detail, a regression analysis revealed that the most significant influence on article choice is exerted by an interaction group of specificity, scope and speaker
knowledge. Moreover, noticeable usage differences are found between participants from the fields humanities and engineering. The results furthermore indicate increased article substitution error rates with partitive indefinites and simple (in)definites, which is unexpected (c.f. Table 3). Reasons explaining these results could either be found in the ongoing grammaticalization process of articles in Mandarin Chinese or the proficiency level of the participants, since the Fluctuation Hypothesis is an explicit learner phenomenon. Similar to previous studies on determiner usage, omission errors are generally low in the data.

<table>
<thead>
<tr>
<th></th>
<th>simple definitiv</th>
<th>partitivity</th>
<th>difficult</th>
</tr>
</thead>
<tbody>
<tr>
<td>correct</td>
<td>81 %</td>
<td>78 %</td>
<td>48 %</td>
</tr>
<tr>
<td>substitution</td>
<td>18 %</td>
<td>22 %</td>
<td>40 %</td>
</tr>
<tr>
<td>omission</td>
<td>1 %</td>
<td>1 %</td>
<td>13 %</td>
</tr>
</tbody>
</table>

Table 3: Results for Simple Definites, Partitivity and Lexically Complex Contexts

Tentatively interpreting the results, this study finds that specificity, in combination with scope and speaker knowledge, has an influence on learners’ article choice, even though article choice in English is solely governed by definiteness. Additionally, different proficiency levels seem to favor different error patterns. However, these findings should not be overgeneralized, as they apply only to the relatively small datasets that control for numerous sociolinguistic variables, such as age and education. In the future, it would therefore be interesting to carry out a larger scale study with regression analysis in mind from the very beginning, as this would allow a more definite answer to the possible influence of the different variables.

References


of the English definite article THE by Turkish speakers. Novitas-ROYAL (research on Youth and Language), 5(2), 183–198.


Simplifying terminology extraction: OneClick Terms
Vít Baisa (Masaryk University, Czech Republic), Jan Michelfeit (Lexical Computing Ltd, Czech Republic) and Ondřej Matuška (Lexical Computing Ltd, Czech Republic)

Despite the fact that methods for terminology extraction represent a heavily studied topic, many industrial applications use rather simplistic solutions to tackle this problem (see Steurs, 2015, for an overview). They generally work with word forms (not lemmas) and word n-grams and their occurrence counts in the analyzed document. This leads to results which are often heavily polluted by items that do not qualify for terms and a lengthy manual cleaning process is needed to produce a good quality list of terminology.

The state-of-the-art approaches can, however, extract terminology lists which require little or no manual cleaning at all. This requires that the source text be lemmatized and tagged for parts of speech which will create the base for applying two complementary principles: unithood and termhood.

Unithood is the quality of a lexical item to qualify for a term in a language. A combination of preposition + verb + preposition will not be considered a valid term structure in most languages while adjective + (optional) adjective + noun will. Language dependent rules referred to as term grammar defining valid word combinations are applied during terminology processing.

Termhood is the quality of a lexical unit to be specific to the domain. Termhood is established by comparing the frequency of the whole (possibly multi-word) lexical unit in the analysed text to the frequency of the same (multi-word) lexical unit in a reference corpus. Best results are achieved with a large general language reference corpora of billions of words.

The procedure outlined above is already fully functional and available in 20 languages through the Sketch Engine (Kilgarriff, 2014). However, the typical user looking for term extraction is a translator or a terminologist without the ambition of using a complete suite of Sketch Engine features and without the time or will to learn a complex system for completing a task which might seem so basic on the superficial level.

With this in mind, a decision was taken to design a single-purpose user-friendly interface to accomplish this task in as few steps as possible. In Sketch Engine, the user has to visit 5 screens and click about 10 times before the list of terminology is produced. A rather unrealistic target of reducing the numbers to 1 click and 1 screen was set.

To extract terminology, the source text has to be converted into a corpus. A series of complex pipelines was meticulously developed to shield the user from the need to control this process, from taking decisions about the best settings and from launching each step of the corpus creating procedure individually. The corpus creation now happens with a simple drag&drop interface.

The supported formats are: TMX, XLIFF (from version 2.0), PDF, DOC, DOCX, HTML and TXT and the ever growing number of supported languages currently includes Chinese (simplified, traditional), Czech, Dutch, English, French, German, Italian, Japanese, Korean, Polish, Portuguese, Russian, Slovenian and Spanish. Support for more languages is continuously developed.
After dropping the file, the upload starts and the user can decide to launch the term extraction immediately or to interfere with the default settings should they wish to do so. If the user prefers the former, then this is the only click required from the user to produce a clean term list.

Since it is reasonable to expect some users to wish to have a certain level of control over the process, an intermediary settings page was introduced. The settings include easy to understand options with glossary, any originally numerical settings have been converted to visual controls. The user can control the following options:

- lemmatized or non-lemmatized list,
- a slider with 5 stops was introduced to control to which extent the algorithm should prefer rare words, i.e. words specific to the focus text, or common words, i.e. words common relatively frequent in general language,
- the number of items to extract, more terms can be loaded from the result screen,
- a minimum frequency of the term in the focus text (default is 1) can be changed to help filter out certain unwanted items and
- two more options (ON by default) can be switched OFF: the term has to contain at least one letter and the term must not contain non-alphanumeric characters. The latter might be useful with the OFF setting to include various product names or model numbers, e.g. CN-9030b, should this be important to the user.
The result screen lists terminology in two columns: single-word and multi-word items. The user can download the lists as a CSV file supported by a vast selection of software. The result screen also gives direct access to 10 most relevant Wikipedia articles for each term and the user can display the term used in context as it appears in the focus text.

<table>
<thead>
<tr>
<th>Single words</th>
<th>Multi words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Azure</td>
<td>virtual machine</td>
</tr>
<tr>
<td>BizTalk</td>
<td>configuration manager</td>
</tr>
<tr>
<td>PerformancePoint</td>
<td>certification authority</td>
</tr>
<tr>
<td>intercompany</td>
<td>distribution point</td>
</tr>
<tr>
<td>Server</td>
<td>currency unit</td>
</tr>
<tr>
<td>Virtualization</td>
<td></td>
</tr>
<tr>
<td>Deployment</td>
<td></td>
</tr>
<tr>
<td>Visual</td>
<td></td>
</tr>
<tr>
<td>Wizard</td>
<td></td>
</tr>
<tr>
<td>Configuration</td>
<td></td>
</tr>
<tr>
<td>pane</td>
<td></td>
</tr>
<tr>
<td>DPM</td>
<td></td>
</tr>
<tr>
<td>Viewer</td>
<td></td>
</tr>
<tr>
<td>Microsoft</td>
<td></td>
</tr>
<tr>
<td>Desktop</td>
<td></td>
</tr>
<tr>
<td>synchronization</td>
<td></td>
</tr>
<tr>
<td>Authentication</td>
<td></td>
</tr>
<tr>
<td>RemoteFX</td>
<td>managed property</td>
</tr>
<tr>
<td>Studio</td>
<td>price variance</td>
</tr>
<tr>
<td>IIS</td>
<td>layout view</td>
</tr>
<tr>
<td>BitLocker</td>
<td>host group</td>
</tr>
</tbody>
</table>

What seemed rather unrealistic at the beginning was achieved. The user can, without any knowledge of NLP, text corpora, tagging or corpus management software, drop a file into the interface, click once and produce a clean list of terminology. The length of the whole process is dependent on the size of the source file and lasts only a couple of seconds for average-sized files. The most significant achievement is the sheer simplicity. The user can operate the system instantly, without the need to take any decisions and without any introductory training. This led to the decision to give the system a name that describes this achievement best: *OneClick Terms*.

**References**


Firth’s (1957: 190) argued that the study of meaning is a permanent interest of scholarship. He (Firth 1957: 192) further suggested the procedure for dealing with meaning, that is, to disperse it into different modes, among which colligation is one salient type. Firth’s notion of colligation is rather complex which means the co-occurrence of multi grammatical choices such as the classes of word and sentence or other similar grammatical categories. The statement of meaning at the syntactic level is in terms of these grammatical categories and of the interrelation of them in colligations. Corpus linguists borrow Firth’s term of colligation but do not apply full colligation in their description of grammatical patterns. Colligation is narrowed down to the co-occurrence of word class with a collocating pair (Sinclair 1991, 2004; Hunston & Francis 2000), mainly because of the limitation of the present descriptive techniques on the syntagmatic axis (Sinclair 2004: 142). However, this simplified colligation is an extremely useful concept in describing the grammatical structure and stating the meaning at the grammatical level. Sinclair (2004) used it in his model of extended unit of meaning, Hunston & Francis (2000) employed it in their Pattern Grammar, and consequently it has been widely accepted by corpus researchers. However, Teubert (2007: 227) argued that “grammatical categories have to be posited on more than just the one level (namely word class)” and proposed “a set of categories and classifications taken from valency theory” (ibid) and this model would supplement Pattern Grammar and “have solutions for the problems of ambiguity and similarity” (ibid: 232). Sinclair (2004: 18) predicted that “valency grammar ... is likely to see an upsurge of interest in the next few years”. However, this is not a revival of interest in the traditional valency grammar, but in capitalizing some categories and classifications in valency grammar to augment pattern grammar.

The privilege of valency is to link local, lexical grammar to general grammar (Teubert 2007: 225). And according to Sinclair (2008), in describing meaning, grammar and lexis cannot be separated. Inspired by these views, Reichardt (2014) proposed the notion of valency sentence patterns and explores the meaning interpretation of the valency patterns, taking the verb CONSIDER as a case study. However, Reichardt’s study has a limitation, namely the narrow scope of empirical data. If our research interest is in the integration of lexis’ valency patterns and meaning, more data are desirable. Hornby (1954, cited in Hunston & Francis 2000: 5) once suggested to deal with together a group of verbs with similar patterns to identify common meaning.

The present research aims to investigate the integration of lexis’ valency
patterns and meaning. To achieve this aim, we apply Reichardt’s system of valency pattern to the description of a group of verbs, namely, consider, believe, feel, find, guess, judge, know, prove, suppose and think, as suggested by Hornby. The corpus used is the BNC. Firstly, 200 randomly-chosen concordance lines are retrieved from the BNC for each verb. Secondly, the verbs are described in terms of valency patterns, and the valency patterns for each verb are listed in descending order of frequency. Thirdly, common valency patterns among all the verbs are identified, and as well as the idiosyncratic patterns. Finally, the association among lexis, valency patterns, and meaning is observed. The major findings are presented as follows.

1. There are some common valency patterns for all the verbs, such as “Sub V Obj-that”, “Sub V Obj-nom”, “Sub V Obj vb-to-inf”, “it BE adj to V Obj-that”, “Sub V Obj-wh”, etc., and some for most of the verbs, such as “Sub V Obj adj”, “Sub V”, “it be V-ed that”, “Sub be V-ed vb-to-inf”, etc. The common valency pattern identified indicates a common sense among all the verbs involved.

2. Each verb has its preference for some typical valency patterns. For example, the verb prove is highly associated with “Sub V adj”, and the verb guess tends to occur very often in “Sub V”. This preference can be used as an indicator of difference among the verbs.

3. Some verbs are repulsed by certain valency patterns which are associated with other verbs. For example, “Sub V Obj adj” is frequently linked with the verbs find, believe, feel, judge, prove, think, but the verbs guess, know, suppose are repulsed by it. In order to validate this finding, we examine the association between the verb and the valency pattern in the whole BNC. The repulsion between a verb and valency pattern is of great value in analyzing the absence of meaning in the verb.

4. Valency patterns can be interpreted in terms of meaning. In other words, sense and structure are interwoven. For example, the verbs in the valency pattern “it BE V-ed that” share a common sense, namely to objectify, albeit in varying degrees, the statement expressed in the “that-clause” following the verb.

5. Some valency patterns are observed to have obvious pragmatic functions which have little to do with the semantic meaning of the verbs involved. For example, the valency pattern “Sub V Obj adj” implies evaluation, and most of the verbs under survey can be used in this valency pattern, no matter what semantic meaning they have. Moreover, the same pragmatic function can be realized by different forms of valency patterns. For example, evaluation can be realized by “Sub V Obj adj”, “Sub V Obj nom”, “It BE adj V Obj-that”, and so on.

The above findings, albeit tentative, indicate the complex association between sense and structure. And these findings also have pedagogic implications, since
students, especially EFL students, hope to learn words with semantic and syntactic restrictions.

References

Ample evidence has shown that language is highly patterned or phraseological (e.g. Sinclair, 1991, 2004; Hunston & Francis, 2000). Construction grammarians hold that language consists of millions of constructions which are form-meaning mappings, conventionalised in the speech community and entrenched as language knowledge in the learner’s mind (Robinson & Ellis, 2008; Ellis, 2013). The widespread recognition of usage-based approach to constructions or phraseologies has made Corpus Linguistics a most viable methodology to scrutinise such frequent constructions as verb-argument constructions (VACs) in learner language.

The present study attempts to examine the use of VACs in oral English production by Chinese EFL learners of different levels of proficiency. Our focus will be on the semantics of the verbal constructions in light of collostructional statistics as well as the comparisons of learner groups proficiency level. The corpus used in this study is SECCCL (Spoken English Corpus of Chinese Learners), i.e. the spoken part of SWECCL (Spoken and Written English Corpus of Chinese Learners). SECCCL is a corpus of one million words in size, which consists of oral production of intermediate learners (Test for English Majors Band 4) and advanced learners (Test for English Majors Band 8), totaling 775, 151 and 286, 309 words respectively. The study focuses on learner’s use of VACs, such as the “V at n construction (e.g., she looked at the picture). Three constructions have been chosen from COBUILD Grammar Patterns 1: Verbs (Francis, Hunston & Manning, 1996) for the present study, which are as follows: V about n, V with n, and V in n.

Firstly, we retrieved the three constructions from the two sub-corpora of SECCCL, and manually checked the concordances to remove the noise. Thus, a list of verbs used in each construction were generated and lemmatised (e.g., walking, walked, walks → WALK), and then the semantic prototypicality of the VACs was measured and analysed according to the collocalar strength of each word with the construction using an R script (<coll.analysis.r>) provided by Gries (2007). The results show that there is no difference between the number of VACs used by learners of different levels of proficiency, but the verbs generated in each construction are not quite the same. With regard to the collostructural strength of each verb in the constructions, we found that the strength of most of the verbs produced by advanced learners is stronger than that of the intermediate learners, and that the order of verbs (ranked according to collostructural strength) varies remarkably between different learner groups. After categorising these verbs according to Levin’s (1993) classification of verb classes, still we found that the collostructural strength of most categories of verbs produced by advanced learners is stronger than that of the intermediate
learners, and the order of verb categories (ranked according to collocational strength) varies between advanced learners and intermediate learners. It is hoped that our findings would shed light on L2 learners’ knowledge of VACs as well as the crosslinguistic influence that impacts verb semantics (cf. Talmy, 1985) of learners’ spoken English.

References


Murderer, mother, slave, or skivvy: XML annotation to enable social actor analysis in a small corpus of English sentencing remarks for women who kill
Amanda Potts (Cardiff University, UK)

Although women make up a minority of homicide convictions in England and Wales, this is a statistically and socially significant minority. Their comparatively minor involvement in such criminal behaviour has resulted in a comparatively small (but growing) body of research engaging with such cases. Much of emerging scholarly work argues that these women are constructed in relation to their deviance from appropriate femininity and are ultimately constructed as ‘mad’, ‘bad’ or ‘victims’, depending on the circumstances surrounding their cases and whether they successfully use any legal defences (see for example: Seal, 2010; Weare, 2013).

This paper utilises corpus linguistic tools on a specialised corpus of courtroom language to explore constructions of women who kill. This group might be considered triply deviant, for: 1) being a small minority group, in 2) committing a very serious crime, which 3) goes against traditional conceptions of appropriate feminine behaviour. Though socio-cultural conceptions of gender should not contribute to legal proceedings, we will demonstrate how judge’s sentencing remarks routinely stereotype women who kill, further ‘othering’ them, removing their agency, and seemingly impacting the sentencing outcome.

This paper illustrates some of the latest work being undertaken at the legal-linguistic interface. Using methods from corpus linguistics, we analysed recurrent (statistically significant) patterns in a collection of sentencing remarks to explore how female defendants are routinely constructed. Analysis is based on a custom-collected corpus of English Crown Courts sentencing remarks, 2012-2015, where a jury has found a woman guilty of unlawful killing. At present, the corpus contains 17 sentencing remarks comprising over 50,000 tokens total. The corpus is analysed in SketchEngine (https://www.sketchengine.co.uk), making use of XML compatibility and WordSketch features.

The small size of the corpus reflects the comparatively small number of women convicted of a homicide offence, but has allowed for full reading of the texts and manual XML annotation (Hardie, 2014) of salient features therein. This approach allowed us to search for all references to women who kill using single query, and to view all instances in context. Based on this, we could indicate frequency of naming strategies and calculate collocates of women who kill, which informed our discussion about the social construction of these participants by judges in sentencing remarks.

Women who kill are referred to 1,809 times in the corpus. Several findings emerged from our initial analysis. Frequency analysis of the textual contents of the XML tags provides insight into common referencing strategies. These women were most frequently referred to directly by the judge using the second person pronoun, but additional functionalization (van Leeuwen, 2008) and nomination strategies construed these homicidal women in terms of familial position, e.g. “you, Keanu’s mother”. The use of such nomination strategies construing these women in terms of familial position also reflects the perceived gendered deviance of women who kill, who have both departed from, and undermined, their roles as the care-givers and
nurturers within society. Annotating all references to a class of social actor also allows for collocation on this group (including even low-frequency instances). Collocations—with concordance analysis—have initially showed a lack of agency associated with women who kill. Verbal processes and mental processes of emotion associate most strongly with women who kill, while active, material processes are much less frequent and are of lower statistical significance. As a result, rather than acknowledging the choice of these women in relation to their homicidal behaviours (which would, in turn, actively challenge existing gender discourse), excuses and explanations are typically invoked when exploring their offending, and the women offenders are routinely constructed as being controlled by mental illness or deviant sexual impulses, or victimised by others.

We sum up by reflecting on the strengths and limitations of working with a very small, extremely specialised corpus of legal language. This work contributes to both legal and linguistic studies. From a corpus linguistic perspective, close reading and manual annotation of features is increasingly rare. From a legal perspective, applying corpus linguistic approaches to legal texts allows for a previously unexplored depth of analysis of these texts to be undertaken. In turn, this allows us to further develop understandings of the discourses and narratives that emerge and thus take steps to combat those that are problematic and damaging.

References


Corpus triangulation: Towards a new methodological framework for translation studies
Sofia Malamatidou (University of Birmingham, UK)

Triangulation, which refers to the combination of multiple (two or more) theories, data sources, methods, or investigators in one study of a single phenomenon, has been employed in a range of disciplines, including the social sciences (Denzin, 1989, 2006; Webb, Campbell, Schwartz, Serchek, & Grove, 1981), nursing and health services research (Begley, 1996; Fotheringham, 2010), education (Altrichter, Posch, & Somekh, 2008; Cohen & Manion, 2011), management (Jack & Raturi, 2006), and applied linguistics (Dörnyei, 2007; Magnan, 2006). The idea of triangulation also found fertile ground in corpus-based translation studies when scholars realised that a combination of different types of corpora is crucial if research in corpus-based translation studies is to move forward. For instance, it is argued that parallel and comparable corpora should not be seen as mutually exclusive, but rather as “complementary sources of cross-linguistic data” (Altenberg & Granger, 2002, p. 9), and that the examination of a monolingual comparable corpus of translated texts needs to be complemented with the comparison of these texts and their source texts, as well as a reference corpus in the languages involved in the comparison (Zanettini, 2000).

Despite the recognition that corpus-based translation research would benefit from the triangulation of corpora, little has been done in the direction of actually employing combined corpora in translation studies. Studies in which some sort of corpus combination has taken place do not foreground the triangulation approach and hardly any treat triangulation in a principled way. This state of affairs is evidence of the fact that corpus-based translation studies, although employing triangulation techniques, is not recognising their full potential. It also raises the important question of whether these studies have employed ‘real’ triangulation techniques, and not an ad hoc collation of different corpus data and, ultimately, of how exactly triangulation is understood within corpus-based translation studies. What is missing is a guided and principled account of how the principles of triangulation can be applied to corpus-based research, which could serve as a model and encourage triangulation to be used more widely but also in a more integrated manner.

This presentation aims to address this gap by providing a clear account of corpus triangulation techniques and demonstrating how these can be applied to empirical data. Following the definition of triangulation in the social sciences (Denzin 1989), where it has been extensively used, corpus triangulation is understood as the combination, in an integrated manner, of multiple (two or more) corpora in one study of a single phenomenon, for example the combination of a comparable monolingual corpus and a parallel bilingual corpus. To avoid the risk of loose interpretation and help refine further how exactly corpora can be combined, the corpus triangulation framework is associated with a corpus typology based on the idea of variables, values and attributes. This typology is introduced in this presentation together with clear examples of how it can facilitate triangulation to avoid further ad hoc attempts towards corpus triangulation. The purpose of corpus
triangulation is to increase knowledge about the linguistic and/or translation phenomenon under investigation, by approaching it from different perspectives and examining various relevant parameters. Thus, by providing corpus-based translation studies with a clear framework for corpus triangulation, which is based on a comprehensive and flexible corpus typology, the quality and rigour of research conducted in the field can be improved, as clarity and objectivity increase. This means that not only can we answer existing questions in translation studies with more confidence, and increase the reliability of research, but that we can also attempt to address new questions, which have been left unanswered until now.

To demonstrate how corpus triangulation can be used in corpus-based translation studies, the methodological framework is applied to a case study focusing on the examination of the language of translation. In particular, corpus triangulation techniques are used for the investigation of the pragmatic factors affecting the use of adversative connectives in Russian translated texts from English. The factors examined are the genre of the text (fiction, children’s fiction, and non-fiction), the audience it addresses (adults vs. children), the influence from target linguistic conventions, and the influence from the source texts. Connectives are considered optional linguistic units since their presence in a text is not a necessary condition for the establishment of a link between two word groups, clauses or sentences. This optionality makes connectives a good candidate for cross-linguistic examination, especially when mediation is involved in the form of translation. Regarding function, adversative connectives (e.g. but, yet, although in English and а, на самом деле, хотя in Russian) signal that the new information will contradict the information already available as in Examples (1) and (2) below.

(1) Yet although they lasted as long as the Soviet Union itself, and although many millions of people passed through them, the true history of the Soviet Union’s concentration camps was, until recently, not at all well known.

(2) Хотя горные снега и подземные воды, на первый взгляд, имеют мало общего, на самом деле они связаны. (Although mountain snow and underground water, at first glance, have little in common, actually they are related)

Texts include translated and non-translated Russian texts, and their English source texts, totalling a 9-million-word corpus. By combining these texts in different ways, four different subcorpora are compiled each associated with a different stage of analysis. The first stage consists of the analysis of a comparable bilingual (English and Russian) subcorpus of non-translated texts to examine how adversative connectives are distributed in Russian and English non-translated texts across different genres. Then, a comparable monolingual (Russian) subcorpus of translated texts is used to examine how adversative connectives are distributed across three translated genres in Russian. The third stage involves the examination of a comparable monolingual (Russian) subcorpus of translated and non-translated texts to investigate whether any observed differences might be related to the linguistic preferences found in respective non-translated Russian genres. Finally, a parallel bilingual (English-Russian) subcorpus of source and target texts is analysed to establish whether the observed differences might be traced back to English. Each stage of analysis provides a partial answer to the research question, and results
from each stage interact, cross-fertilize ideas and provide insights. It is only by combining the results from all three stages that a clear conclusion can be reached. A unique aspect of the corpus methodology, which is evidence of the integrated approach adopted, is that the first two stages of analysis rely more heavily on generating corpus results and examining the distribution of connectives in each subcorpus, while the last two stages of analysis rely more on comparing and contrasting the results generated during the first two stages of analysis.

The corpus triangulation techniques employed in the case study allow for meaningful comparisons to be made across different types of texts, in an effort to acquire a comprehensive understanding of the phenomenon of adversative connectives and how these are employed in translation. Results suggest that a complex interplay of factors affects the use of adversative connectives in Russian translation from English and these can be related to the influence from existing target language linguistic conventions, the source language interference, and audience, or other genre-specific, considerations. In particular, translated Russian fiction is very close to non-translated fiction in the same language, and the influence of the existing linguistic conventions in this genre is much higher than any other influence. Regarding children’s fiction, translated Russian texts make a much higher use of adversative connectives than both non-translated Russian and English children’s fiction. This could be considered an indication that considerations of comprehensibility have preoccupied translators, who wanted to create more accessible texts for children in particular. Finally, regarding non-fiction, translated texts seem to be somewhere in between Russian non-translated and English non-fiction, suggesting that some influence from English source texts might have been exerted. Had corpus triangulation techniques not been employed, it would not have been possible to examine the way in which these different factors interact and, ultimately, affect different genres to a different degree and level.

References


The study of syntactic complexity has attracted considerable attention in second language (L2) writing research in recent years, revealing the range of grammatical resources employed by learners in their text construction and the role these resources play in L2 writing development (e.g., Ortega, 2003; Ai & Lu, 2013; Wood & Struc, 2013; Kim, 2014; Scontras et al., 2014; Lu & Ai, 2015; MacDonald, Montag & Gennari, 2015; Mazgutova & Kormos, 2015; Vyatkina, Hirschmann & Golcher, 2015; Biber & Gray, 2016). Most studies conducted in this area have been cross-sectional in design (cf. Wood & Struc, 2013; Vyatkina et al., 2015) and often involve comparisons with a target language norm. Several early studies, which focused on a number of specific syntactic measures, have provided some initial findings which offer useful longitudinal insights (see, e.g., Larsen-Freeman, 1983; Arnaud, 1992; Kern & Schultz, 1992; Casanave, 1994; Ishikawa, 1995). More longitudinal research, however, is required to make sense of how changes in the employment of a wide range of grammatical resources might take place over time, which is essential to a fuller understanding of what development may mean in the study of L2 writing.

This paper reports on a longitudinal corpus study that considered 14 syntactic complexity measures in L2 texts to determine whether and to what extent changes may be observed in the writing of a group of 124 secondary school students at four points in time over a 24-month period. All the texts produced by the students were based on a common narrative task. The results of the 14 measures are generated using L2 Syntactic Complexity Analyzer (Lu, 2010), which examined such constructs as mean length of sentence (MLS), mean length of T-unit (MLT), mean length of clause (MLC), clauses per sentence (C/S), verb phrases per T-unit (VP/T), clauses per T-unit (C/T), dependent clauses per clause (DC/C), dependent clauses per T-unit (DC/T), T-units per sentence (T/S), complex T-unit ratio (CT/T), coordinate phrases per T-unit (CP/T), complex nominals per T-unit (CN/T), and complex nominals per clause (CN/C). The scores from the L2 Syntactic Complexity Analyzer were analyzed using repeated measures ANOVA to ascertain if there are statistical differences of the 14 indices as observed across four points in time the writing of the 124 students.

The findings of the study indicate a statistically significant change in the use of syntactic structures based on 10 out of the 14 measures over time (i.e., MLS, MLT, DC/C, DC/T, CP/C, CN/C, C/S, VP/T, CN/T and C/T). More dependent clauses per clauses and per T-unit, complex nominal per clauses,
mean length of T-unit, clauses per sentences and complex nominal per T-unit were observed to be used over time in the student texts. The use of coordinate phrases per clauses, mean length of sentences, verb phrases per T-unit and clauses per T-unit, on the other hand, have significantly decreased over time. Interestingly, while a high employment of dependent clauses has been noted to be a feature of spoken language (e.g., Halliday, 1989; Biber et al, 2011), this category of clauses has been found to be used more frequently over time in this study (see also Pallotti & Ferrari (2008) for similar observations). A heavy use of subordinations also suggests high causative relations in writings (Ryshina-Pankova, 2015), indicating an increased emphasis on such relations in the narrative texts. Equally interesting, as time goes by, the use of complex nominals was found to have significantly increased, which opens up questions about the use of complex nominals in genres other than academic writing (cf. Biber & Gray, 2016). Furthermore, the findings show that while changes as measured in certain constructs can be observed within six months, changes in other syntactic features may only be observable over a full period of 24 months, indicating an organic, developmental process at play (e.g., Chau, 2015; Larsen-Freeman, 2015).

The paper concludes by offering some cautionary notes. While the observed increase in complexity may imply that there is evidence of more sophisticated language use over time in the writing of the students, it should be remembered that “more complex does not necessarily mean better” (Ortega, 2003). Development in L2 writing essentially entails a complex interplay between one’s deployment of linguistic resources and what is socially valued as ‘good’ writing (see, e.g., Skehan & Foster, 2007; Ortega, 2015). As Pallotti (2009, p. 14) observes, the ability for learners to put language into use involves not only syntactic complexification; it “also entails the development of discourse and sociolinguistic repertoires that the language user can adapt appropriately to particular communication demands”. Further implications for future research into syntactic complexity in L2 writing are considered.

References


Modal verbs are frequently used in every day communication. In previous studies, modal verbs were studied from different perspectives, such as the formal linguistics, semantics or pragmatics (Thomas, 1983; Leech 1983, 2003; Jacqueline & Kenneth, 1997; Rappen, 2002; Neil, 2009; Peters, 2013). Halliday (1994) proposed that VALUE is an important factor to represent the variability in modality. Halliday and Hasan (1989) divided modal verbs in terms of their pragmatic values. High value modals include must, ought to, need, and have to; intermediate value modals include will, would, shall, should; and low value modals include may, might, can, could. Different groups of modality are related with different politeness degree of the speech. Modal verbs of high value indicate an impolite speech, which is liable to cause the reader/listener’s disfavor, whereas low value modals suggest a most polite use of language. In L2 learning, modal verbs appear as a challenge for learners for its various communication values and functions. With the development of computer technology and corpus linguistics, studies on modal uses by L2 learners can be based on corpus data analysis.

The present study attempts to explore the different uses of modal verbs by English L2 learners and native speakers. It compares two writing corpora, a Hong Kong University Students Writing Corpus and the British University Students Writing Part of LOCNESS (Granger, 1995). The following research questions are raised in the study: 1) Do English learners prefer to use modal verbs of high value? 2) Do English learners use more modal verbs than the native speakers? 3) Are there any different or similar tendencies in terms of modal verbs use between English learners and native speakers? Concordancer in Concapp V4 has been used to search for the modal verbs, and their frequencies were compared for different research questions.

For research question 1, it was found that among the four modals with high value, learners used “need” and “have to” twice of their British counterparts, and “must” 1.5 times of the native speakers. As for the intermediate and low value modals, learners used more intermediate modals than native speakers do, for example, “should” was used three times by learners. In addition, British students tended to use “would” and “could” much more frequently than their non-native counterparts.

For research question 2, it was found that for the overall frequency of all the modal verbs, learners used more modals (1.5 times) than the native speakers. And for research question 3, it was revealed that first, there was a similar tendency between the two groups----both English learners and native speakers tend to use “can” and “will” as the most frequent modals, and both of them used “might” “ought to” and “shall” less frequently. There was a significant difference between learners and native speakers in terms of the use of “could”. British students used “could” frequently, whereas learners seldom used this word.

This paper then explored the reasons for the modal use differences between L2 learners and native speakers. First, Learners tended to use high value modals which indicated an absence of reader/audience awareness. It seemed that learners were not well informed of the pragmatic meaning of modal verbs.
Second, learners used not only more high value modals, but also more low value modals than native speakers. In brief, learners overused modal verbs in their writings. This suggested that in most cases, learners used modals only to express the possibility or necessity of certain propositions, with no concern about the pragmatic functions of modal verbs. For the overuse of modals, we can explain the phenomenon in the following aspects:

1) Concerning its semantic and syntactic features, modal verbs do not need inflections to indicate tense, person, voice etc. In other words, it is a simple form to be adopted by learners. And there is no complex syntactic rule for modals. As such, L2 learners prefer to use modal verbs even when it is not necessary.

2) Modals have a wide range of communication functions. Take “can” as the example. “Can” indicates different functions in multiple contexts. That is the reason why both non-natives and natives used “can” most frequently among all the modals. For natives, they used “can” frequently to fit different types of contexts, whereas the overuse of modals by learners suggested that learners did not understand the minute differences among modals, and they used one to fit all contexts.

3) Learners and natives have similar preference for modals use. They preferred to use “can, will and should/would”, and used “might, ought to, shall” as the least frequent ones. “Might” is called the mildest modal to show the highest degree of politeness (Leech, 1983). At the same time, the highest degree of politeness indicates the biggest distance between interlocutors. In daily writing and conversation, “Might” is used less frequently than other modals. When “might” is used to indicate possibilities, it shows the least possibility, which reflects uncertainty in writing and speaking. As such, it was used least by both learners and natives.

Another post hoc finding is that “could” was used distinctively by learners and natives. Among all the modals, “could” was used less frequently by learners than most of other modals, whereas natives used “could” more frequently than other modal verbs. “Could” indicates not only past tense, but politeness or uncertainty. “Could” goes between “might” and “can” when pragmatic connotations are concerned. Although natives have no difficulty to distinguish the three modals “might, could, can”, it is a challenge for learners to use this modal verb properly.

This study suggested that learners need to be informed of the following two types of information about modal verbs: a) the pragmatic connotations of modals; b) the detailed differences between different modal verbs and the context differences. The lack of the above information results in a poor performance in modal verbs use by L2 learners. A series of L2 teaching suggestions have been proposed for the instruction and acquisition of English modal verbs.