ICAME42 | tu

crossing boundaries through corpora

August 18-21, 2021 | TU Dortmund University

Book of Abstracts

Table of Contents

1) Plenaries 9

Nagy, Naomi
Is a heritage variety just a regional variety spoken outside the national boundary? 9

Schweinberger, Martin (University of Queensland/Arctic University of Norway) 10
Corpus Linguistics, Language Data Science, and Computational Linguistics – building bridges or splitting apart? 10

Suárez-Gómez, Cristina 11
From multilingual to monolingual Gibraltar: the emergence of a local identity and the spread of Gibraltar English 11

Szmrecsanyi, Benedikt (KU Leuven)
What corpora can tell us about the link between grammatical variation and language complexity 12

2) Workshops

Workshop 1: Corpus pitfalls: dealing with messy data (and other traps for the unwary)13
Convenors: Mark Kaunisto (Tampere University), Marco Schilk (Universität Hildesheim), Jukka Tyrkkö (Linnæus University)

Callies, Marcus (Universität Bremen)
Challenges in the annotation and analysis of learner corpora 15

Egbert, Jesse (Northern Arizona University), Tove Larsson (Uppsala University), and Douglas Biber (Northern Arizona University)
Research design in quantitative corpus linguistics: Critical reflections and suggested improvements 17

Hartmann, Stefan (University of Bamberg)
Open Corpus Linguistics: Overcoming Rissanen's problems (and others) with open data 18

Hiltunen, Turo (University of Helsinki)
Issues in using *British Library Newspapers* as a corpus 19

Kaunisto, Mark (Tampere University)
Proper names as potentially problematic items in corpora 20

Miletic, Filip, Anne Przewozny-Desriaux and Ludovic Tanguy
Modeling fine-grained sociolinguistic variation: the promises and pitfalls of Twitter corpora and neural word embeddings 21

Sundberg, Daniel (Linnaeus University)
Corpus Categories: What and for whom? When special corpora meet general corpora in comparative studies in literature 23

Tyrkkö, Jukka (Linnaeus University) and Sophie Raineri (Paris Nanterre University)
Empirical perspectives on the reliability and accuracy of collaborative pragmatic annotation 25

Vetter, Fabian (University of Bamberg)
Register variation & comparability in parallel corpora 27

Workshop 2: Crossing Language and Discipline Boundaries
Convenors: Anna Čermáková (University of Cambridge, UK / Charles University, Czech Republic), Hilde Hasselgård (University of Oslo, Norway), Markéta Malá (Charles University, Czech Republic), Denisa Šebestová (Charles University, Czech Republic) 29

Aijmer, Karin
The contrastive workshop - the first ten years 30

Bourgoin, Charlotte, Kristin Davidse, Karen Lahousse (KU Leuven)
A framework for cross-linguistic study of the information structure of English it- and French c'est-clefts 31

Čermáková, Anna and Lenka Fárová
Reporting verbs in English, Czech and Finnish: *said* and beyond 33

Egan, Thomas (Inland Norway University of Applied Sciences)
The dative alternation in English and Norwegian: verbs of SENDING, BRINGING, LENDING and SELLING 35

Hasselgård, Hilde (University of Oslo)
Periphrastic genitive constructions in English and Norwegian 37

Martinková, Michaela, Markéta Janebová, Palacký University, Olomouc
Coronavirus pandemic across languages and waves: Evidence from newspaper-based corpora of English and Czech (WiP) 39

Oksefjell Ebeling, Signe (University of Oslo)
Seeing through languages and registers: A closer look at the cognates *see* and *se* 41

Rabadán, Rosa, Camino Gutiérrez-Lanza
Corpus-based contrast in audiovisual customization (w-i-p) 42

Ström Herold, Jenny and Magnus Levin (Linnaeus University)
From dashes to dashes? – a contrastive corpus study of dashes in English, German and Swedish 44

Workshop 3: Exploring Powerful Tools to Ensure Robust and Reproducible Results in Corpus Linguistics
Convenors: Martin Schweinberger (The University of Queensland, Australia), Joseph Flanaghan (University of Helsinki), Gerold Schneider (University of Zurich) 46

Haugh, Michael (The University of Queensland) and Simon Musgrave (Monash University)
Looking for a good laugh: A combinatorial approach to identifying pragmatic phenomena in spoken corpora 47

Schneider, Gerold (University of Zürich)
Text Crunching Center (TCC): Data-driven methods for linguists, social science and digital humanities 49

Schweinberger, Martin (UQ, UiT)
The Language Technology and Data Analysis Laboratory (LADAL) - building computational humanities infrastructures: experiences, problems, and potentials 50

Wallis, Sean
How do we understand significance? 51

Weihs, Claus and Sarah Buschfeld
Variable importance in ensembles from undersampling: generalizability meets robustness 52

Workshop 4: Rescoping the theory and methodology of linguistic epicenters in World Englishes

Convenors : Pam Peters (Macquarie University, Sydney, Australia),Tobias Bernaisch (Justus Liebig University Gießen, Germany) 53

Götz-Lehmann, Sandra
Caught between epicenters? Tracing influences on English in Nepal 55

Gries, Stefan Th., Benedikt Heller & Tobias Bernaisch
Methods for detecting possible traces of epicentral influence: synchronic and diachronic perspectives 56

Hundt, Marianne, Dirk Pijpops and Laetitia van Driessche
Approaching epicentral influence with agent-based modelling: simulating factors in the choice between verb complementation patterns in speakers of Philippine and Indian English 58

Korhonen, Minna and Adam Smith
Parliamentary Hansard as a focus for regional variance and epicentral influence in Australia, New Zealand and Papua New Guinea. 60

La Peruta, Roberta
BE THAT AS IT MAY: Epicenters of cross-border dis/similarities in the use of the subjunctive In Canadian vs. American English 61

Mair, Christian (Freiburg)
Global Jamaican: Migration, media and the creation of an informal epicentre 63

Meierkord, Christiane (Ruhr-University of Bochum)
Still in awe of the colonial variety? Post-protectorate Uganda as a challenge to current models of the influence between Englishes 64

Peters, Pam, Adam Smith and Minna Korhonen
"A diachronic study of *-ed/-t* inflectional variation in Australian and New Zealand English: Epicentral influence under the radar?" 66

Schneider, Edgar W. (Universität Regensburg)

Parameters of epicentral status: requirements and traces 68

Workshop 5: To boldly go: Corpus approaches to the language of Science Fiction
Convenors: Claudia Lange (TU Dresden) Sofia Rüdiger (University of Bayreuth) 70

Behrens, Tanja & Dr. Anke Schulz (Universität Bremen)
Where no woman has gone before – five decades on the *Enterprise* 71

Blombach, Andreas, Thomas Proisl, Stefan Evert, Philipp Heinrich, Natalie Dykes
Into the *Perryverse*: A CL Journey to the Realm of Lexical Complexity 72

Eberl, Martin (LMU München)
How "alien" are fictional languages? Approaching 'estrangement' in constructed languages 73

Gee, Matt (Birmingham City University)
"there was much new to grok": A semantic analysis of word creation in science fiction. 74

Laliberté, Catherine, Melanie Keller and Diana Wengler (LMU Munich)
*"*So I trucked out to the border, learned to say *ain't*, came to find work*"*: the sociolinguistics of *Firefly* 75

Oakes, Elizabeth (University of Helsinki)
Representing Altered Consciousness in American New Wave Science Fiction: A Corpus-Stylistic Analysis 76

Ronan, Patricia (TU Dortmund) and Gerold Schneider (University of Zürich)
To boldly go – whence and where? 77

Yurchenko, Asya (Dresden University of Technology)
Quantifying the language of estrangement in science fiction using unique lexical compounds as a measure 78

3) Papers at the main conference

Appleton, Stephen
Using corpus techniques to explore lexical change in parliamentary discourse 79

Balfour, James
FREED TO KILL: a corpus-driven analysis of responsibility in British newspaper articles reporting on schizophrenia 80

Bednarek, Monika
Mediated Australian Aboriginal English: A corpus linguistic study 81

Blake, John
Visualizing Rhetorical Moves in a Corpus of Scientific Research 82

Bernaisch, Tobias and Julia Degenhardt
Apologies in Indian and Sri Lankan English 83

Bernaisch, Tobias and Nina Funke

Intensifiers and downtoners in South Asian Englishes: Corpus-based perspectives on Indian and Sri Lankan English 85

Burgeois, Samuel
On the evolution of parentheticals in journalistic English 86

Brunner, Thomas
We don't play politics with it. The *play*-X-*with*-Y Construction in Varieties of English 87

Busse, Beatrix, Ingo Kleiber, Sophie Du Bois and Nina Dumrukcic
Crossing the boundary of time: Retraining modern NLP models for specialized historical corpus data 88

Bychkovska, Tetyana
Effects of Explicit Instruction on Noun Phrase Production in L2 Undergraduate Writing 89

Castro Chao, Noelia
Minor complementisers in the temporal domain: The case of English *till* and *until* 90

Collins, Luke
Examining “doctor talk” in a corpus of Emergency Department (ED) interactions 91

Correia Saavedra, David, Jennifer Rains and Martin Hilpert
English back-clippings and fore-clippings: What can we learn from corpus data that the dictionaries don't tell us? 92

Crosthwaite, Peter
Integrating corpus use into trainee EFL teachers’ lesson planning: Building corpus- focused TPACK for the future. 93

Dayter, Daria and Thomas C. Messerli
Persuasive register in Change My View subreddit 94

Del Fante, Dario
Figurative language and pandemics: Spanish Flu and Covid-19 in Us newspapers. A case-study. 95

Denison, David, Nuria Yáñez-Bouza, Tino Oudesluijs, Cassandra Ulph and Christine Wallis
Transcribing and Editing *The Mary Hamilton Papers (c. 1750 – c.1820)* 96

Ebeling, Jarle
120 years of reporting clauses: stability or change? 97

Facchinetti, Roberta, Silvia Cavalieri and Sara Corrizzato
Compiling a spoken corpus of journalistic interviews: from transcription to pragmatic annotation 98

Franceschi, Valeria
China for the international traveler. A Corpus-based analysis of early travel guidebooks to China 99

Garretson, Gregory and Rachele De Felice

Detecting and analysing problem-oriented language in an email corpus 100

Gilquin, Gaëtanelle and Lea Meriläinen
Constrained communication in EFL and ESL: The case of embedded inversion 101

Hannaford, Ewan
Interpreting Illness: A corpus linguistic investigation of represenations of mental and physical illness in the UK and US press (1995-2017) 102

Hannader, Helena
Working towards a gold standard in writing revision analysis 103

Hober, Nicole
Hybrids compounds in World Englishes 104

Iyeiri, Yoko
Exploring Benjamin Franklin’s English: A Case of Intra-Writer Variation 105

Kaunisto, Mark and Paula Rautionaho
Was/were variation with plural subjects in recent British English – towards standard uses? 106

Kehoe, Andrew, Matt Gee and Antoinette Renouf
A collocational approach for explaining changes in word frequency over time 107

Kemble, Melissa
Female athletes and masculine sports: A corpus analysis of patriarchal discourses in the Australian print media 108

Klavan, Jane and Sandra-Leele Toom
The use of phrasal verbs by Estonian-speaking EFL learners 109

Kohn, Birgit
When ‘very crazy’ is not crazy enough: Creative ADJ-intensifying constructions 110

Larsson, Tove , Luke Plonsky and Gregory R. Hancock
Using structural equation modeling to accommodate the multivariate nature of language: The case of syntactic complexity 111

Leone, Ljubica (Lancaster University)
Linguistic changes and structural interferences: Phrasal-prepositional verbs in the Late Modern English period 112

Lensch, Anke
Looking intoing showy-offy stayer-onner-for-nowers. A corpus-based study into English morpho-syntax. 113

Leuckert, Sven and Asya Yurchenko
The Plurality of Indian English(es): Syntactic Variation in the *Corpus of Regional Indian Newspaper English* (CORINE) 115

Li, Zeyu , Ulrike Gut and Philipp Meer

Patterns of co-variation in Standard Scottish English: The case of /r/, /ʍ/-/w/, and the NURSE merger 116

Mahler, Hanna
The interpretative progressive: A case of grammatical constructionalisation? 119

Matsumoto, Noriko
The Ongoing Change from the *Remain-UnVed* Sequence to the *Go-UnVed* Sequence 120

Milička, Jiři and Denisa Šebestová
Human Friendly Corpus Query Language 121

Moessner, Lilo
Modality and the English subjunctive. A diachronic study 122

Pitzle, Marie-Luise and Ruth Osimk-Teasdale
Introducing VOICE Online 3.0 – A new web tool for research on spoken ELF interactions 123

Puga, Karin, Kathrin Kircili and Sandra Götz-Lehmann
Syntactic segmentation of spoken corpus data: What prosody can contribute 124

Ranaweera, Mahishi
A corpus-based study of pragmatic markers in the spoken discourse of women in Sri Lankan English 125

Ren, Haoshan
Abstract nouns in Chinese-learner English argumentative writing 126

Ren, Haoshan and Viviana Cortes
Functional language in academic lectures – an investigation of corpus linguistic findings nad implication for language assessment 127

Romasanta, Raquel P.
Geographical location as predictor of language variation in postcolonial Englishes 128

Rodríguez-Puente, Paula
Suffix competition across registers: On the development of *-ity* and *-ness* in the Late Modern English period 129

Römer, Ute
“I can’t sing and dance but I can cook”: How do constructions with modal verbs develop in second language learners of English? 130

Roomäe, Kärt
Deciphering Spoken American English: A Construction Grammar Approach to the *how-x* Construction in MICASE 131

Säily, Tanja, Martin Hilpert and Jukka Suomela
New approaches to investigating change in derivational productivity 132

Šaldová, Pavlína

Postpositive adjectives and participles: crossing the boundary 134

Savchenko, Denys
Adjective intensification in the Spoken language of Estonian learners of English 135

Julia Schlüter
The uptake of corpus evidence on World Englishes by ELT professionals 136

Schneider, Gerold
Digital Dickens: an automatic content analysis of Charles Dickens' novels 137

Sugiura, Masatoshi , Daisuke Abe and Yoshito Nishimura
L1/L2 Differences in Terms of Mean Syntactic Distance (MSD) found in Learner Corpus Data: Some Evidence for the Shallow Structure Hypothesis 138

Sun, Shuyi and Peter Crosthwaite
A corpus-based cross-disciplinary study of *negation* in Ph.D. theses' limitations sections 139

Tagliamonte, Sali and Jeremy Needle
Honest to Pete! Honesty expressions in contemporary North American English 140

Gries, Stefan Th.
A new approach to (key) keywords analysis: augmenting frequency-based keywords using dispersion 141

Schneider, Edgar W.
Lexicosemantic variability in World Englishes: a case study of prospective verbs 142

Weckermann, Michelle
Prepositions in a cognitive linguistics framework: A cross-linguistic comparison of semantic networks between German and English prepositions 143

Weihs, Claus and Sarah Buschfeld
RePrInDT: Resampling in unbalanced corpora 144

Weilinghoff, Andreas
Pushing transciption work to the next level - Using ASR and LaBB-CAT for linguistic studies 145

1) Plenaries

Nagy, Naomi

Is a heritage variety just a regional variety spoken outside the national boundary?

Naomi.Nagy@UToronto.CA

Comparative variationist sociolinguistics examines variable aspects of language, or "different ways of saying the same thing," by simultaneously considering multiple probabilistic factors that impact speakers' choices. These include inter-speaker (social) and intra-speaker (linguistic context) predictors, and account for community members' shared grammars, as well as how we use stochastic information to recognize speakers' group memberships (*cf*. Labov et al., 2011). This approach has been applied to many languages, but most ensuing generalizations are based on studies of large well-documented varieties. In this talk, we will navigate the boundary between (analysis of) standard homeland varieties and less-standardized heritage varieties.

For every variable element of a language, a *Probability Matrix* must be acquired by speakers, containing probabilistic information about when each form is (more) appropriate. Sociolinguists model such matrices through multivariate regression analyses that reveal significant predictors (and levels within each predictor). One way to understand differences between the language varieties used by homeland and heritage[1] speakers is to ask how a Probability Matrix compares between Heritage and Homeland speakers. (How) can these fairly be compared? Although there have been some proposals tendered, the variationist field lacks a robust comparative methodology to determine how/if varieties differ. In this talk, I focus on one weakness: different-sized samples are often compared, as it can be harder to find/build a large corpus of a small, under-documented language. This difference in sample size implicates different levels of statistical significance even when the two populations' patterns are identical.

I illustrate one solution through comparison of variable patterns in Heritage and Homeland Cantonese. Revising analyses conducted previously of two morphosyntactic variables: *prodrop* and *classifiers* (Nagy, 2015; Nagy & Lo, 2019) and applying a bootstrap procedure to mitigate issues associated with unequal-sized datasets frequent in studies of minoritized varieties, I offer a reproducible comparison method (excerpting from Nagy & Gadanidis *fc*). From these analyses, we learn that heritage and homeland grammars' degrees of complexity are similar: their Probability Matrices are the same size. This approach allows us to consider the complexity of the decision-making process the speakers apply in selecting among forms. As one might expect, heritage and homeland speakers are capable of equally complex processes. This adds another report from the Heritage Language Variation and Change in Toronto Project that finds little difference between Homeland and Heritage varieties of 10 languages spoken in Toronto, when applying corpus-based rather than experimental methods.

References

Labov, W., S. Ash, M. Ravindranath, T. Weldon, M. Baranowski & N. Nagy. 2011. Properties of the sociolinguistic monitor. *Journal of Sociolinguistics* 15.4:431-63. https://doi.org/10.1111/j.1467-9841.2011.00504.x

Nagy, N. 2015. A sociolinguistic view of null subjects and VOT in Toronto heritage languages. *Lingua* 164:309-27. https://doi.org/10.1016/j.lingua.2014.04.012

Nagy, N. & T. Gadanidis. *fc*. Heritage language variation and change – How complex is it? *Heritage Language Journal*.

Nagy, N., & Lo, S. 2019. Classifier use in Heritage and Hong Kong Cantonese. *Asia-Pacific Language Variation* 5(1):84-108. https://doi.org/10.1075/aplv.17001.na

[1] A heritage language is one that is not the broader community's majority language; a homeland language is.

Schweinberger, Martin (University of Queensland/Arctic University of Norway)

Corpus Linguistics, Language Data Science, and Computational Linguistics – building bridges or splitting apart?

Computation has doubtlessly revolutionized our world and has become an integral part of all domains of everyday life – from research and finding information over ordering food and booking travels to dating and entertainment – and all this within merely one or two generations. The "digital" has impacted practices in corpus linguistics and has been offering novel pathways for research as we are now able to make use of vast amounts of electronically available data and complex modelling of language and linguistic behaviour (Anthony, 2020). The integration of computation and quantitative statistical means to understand linguistic processes has indeed allowed us to gain important insights into how language changes and cognitive processes underlying language use. It has further resulted in language data- based applications (search engines, machine translation, speech recognition, etc.) that would have been hard to imagine just two generations ago. When it comes to linguistic theory-building, especially the role of frequency – and thus usage-based approaches to understanding cognitive processes – have been rapidly gaining ground. This too would not have been possible without being able to scrutinize large amounts of natural language data.

Yet, this computational revolution has also sparked conflict and has been criticised – leading to arguments that corpus linguistics has become too methods- and data-focused while neglecting the more traditional focus on "linguistic description" (Larsson, Egbert & Biber, in press). This scepticism has raised awareness of potential problems and culminated in discussions about the appropriate use of digital methods – ultimately, the varying responses to linguistics becoming more data-sciency pose questions about the future direction of our field: should or will corpus linguistics in the end remain part of the humanities or could it shift towards engineering?

This talk intends to further this discussion and to address relevant questions relating to the relationship between corpus linguistics, language data science, and computational linguistics using selected case- studies and examples of emerging infrastructure projects. It argues for the need of training and infrastructures in computational skills and the integration of basic programming in linguistics programs, and it discusses how the computational revolution will continue to impact our field.

References:

Anthony Laurence. (2020) Programming for Corpus Linguistics. In Magali Paquot and Stefan Thomas Gries (eds.), *A Practical Handbook of Corpus Linguistics*, 181-207. Springer.

Larsson, Tove, Egbert, Jesse, & Biber, Doug. (in press). On the status of statistical reporting versus linguistic description in corpus linguistics: A ten-year perspective. *Corpora* 17(1).

Suárez-Gómez, Cristina

From multilingual to monolingual Gibraltar: the emergence of a local identity and the spread of Gibraltar English

The presentation is divided into two main parts. The first deals with the emergence and development of English in Gibraltar, a small British Overseas Territory where Spanish, English, Llanito and other languages have long coexisted, in a state of diglossia. I will focus specifically on the 20th and 21st-century socio-historical events that have shaped the current linguistic landscape, in which Gibraltarians, by adopting English as their mother tongue and identity marker, are becoming progressively monolingual. To this end, I apply the Extra- and Intra-Territorial Forces Model (EIF Model, Buschfeld and Kautzsch 2017, 2020), since it allows one to capture the complex interplay of extra-territorial influences on this unusual non-postcolonial setting. Within intra-territorial forces, language attitudes will take centre stage. I will discuss the results of a sociolinguistic study which measures speakers' differential attitudes towards the coexisting languages in Gibraltar. These bring to light the underlying complexity and current speed of change in the linguistic ecology of the territory. The second part analyses a selection of linguistic features of Gibraltar English (GibE) which illustrate its status as an independent nativized variety; drawing on the Gibraltar component of the *International Corpus of English* (ICE-GBR), I will focus specifically on a selection of morphosyntactic features. This will also allow me to briefly report on methodological issues of the compilation of the corpus and, most of all, on the ongoing challenges that we face in Gibraltar, given the low number of speakers and the closed community they form.

References:

Buschfeld, Sarah and Alexander Kautzsch. 2017. Towards an integrated approach to postcolonial and non-postcolonial settings. *World Englishes* 36.1: 104-126.

Buschfeld, Sarah and Alexander Kautzsch. 2020. *Modelling World Englishes. A Joint Approach to Postcolonial and Non-Postcolonial Varieties*. Edinburgh: Edinburgh University Press.

Szmrecsanyi, Benedikt (KU Leuven)

What corpora can tell us about the link between grammatical variation and language complexity

In this presentation, I investigate the relative complexity (i.e., difficulty, see Miestamo 2009) incurred by having to choose between competing grammatical variants. While variational linguists provide overwhelming evidence for the existence, ubiquity, and systematicity of variable patterns — or "alternate ways of saying 'the same' thing" (Labov, 1972: 188), as in *Tom picked up the book* versus *Tom picked the book up* — there are still language mavens and theoretical linguists who dismiss or deplore variability as a matter of doctrine or explain it away (erroneously) as noise or negligible. Nonetheless, the assumption that grammatical variation *could* create undue complexity for language users is not entirely unreasonable. The idea that grammatical variation might burden language production deserves scrutiny not primarily because language users are forced to make grammatical choices — after all, using language *always* entails plenty of choice-making — but additionally because grammatical variation (as opposed to e.g., lexical variation) is typically conditioned probabilistically by any number of contextual constraints (constituent length, animacy, information status, etc.). Even before language users can make a choice as a function of the naturalness of a grammatical variant in a specific linguistic context, they need to check that linguistic context for the various constraints that regulate the variation at hand. It follows that this extra cognitive work must increase cognitive load. Or does it?

Against this backdrop, I report on a study that explores the link between production difficulty and grammatical variability using a corpus-based research design. The idea is that if isomorphism à la Haiman (1985) and No Synonymy à la Goldberg (1995) are design features of human languages, then variation — to the extent that it exists — should be suboptimal. Suboptimality, in turn, should be measurable by quantifying the extent to which variation contexts attract production difficulties.

Contrary to expectation, analysis based on a sub-sample of the Switchboard Corpus of American English (285 transcripts, 34 speakers) shows that the presence of variable contexts does not positively correlate with two metrics of production difficulty, namely filled pauses (*um* and *uh*) and unfilled pauses (speech planning time). When 20 morphosyntactic variables are considered collectively (*N* = 6,268), there is no positive effect. In other words, variable contexts do not correlate with measurable production difficulties. These results challenge the view that grammatical variability is somehow sub-optimal for speakers.

I will conclude by speculating that the putative difficulties introduced by optionality in syntactic structure or morphological realization are offset by a number of counterbalancing benefits of the flexibility to express the same grammatical concept using more than one form.

References:

Goldberg, Adele E. 2003. Constructions: a new theoretical approach to language. *Trends in Cognitive Sciences* 7(5). 219–224. https://doi.org/10/fc25ng.

Haiman, John. 1980. The Iconicity of Grammar: Isomorphism and Motivation. *Language* 56(3). 515. https://doi.org/10.2307/414448.

Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia: University of Philadelphia Press.

Miestamo, Matti. 2009. Implicational hierarchies and grammatical complexity. In Geoffrey Sampson, David Gil & Peter Trudgill (eds.), *Language Complexity as an Evolving Variable*, 80–97. Oxford: Oxford University Press.

2) Workshops

Workshop 1: Corpus pitfalls: dealing with messy data (and other traps for the unwary)

Convenors: Mark Kaunisto (Tampere University), Marco Schilk (Universität Hildesheim), Jukka Tyrkkö (Linnæus University)

In a short but important paper published thirty years ago in ICAME Journal, Rissanen (1989) identified and discussed three potential problems that the use of corpora may present to unwitting scholars not necessarily closely familiar with the contents, structure and overall character of the corpora that they set out to examine. As solutions to the issues raised, Rissanen promoted the compilation of larger and more representative corpora, as well as getting acquainted with the data itself. Today we have access to corpora which are massive, compared to the ones available thirty years ago, which may widely improve representativeness. At the same time, paradoxically, these increases in corpus size make it considerably less likely that corpus users will have a clear understanding of the characteristics of many of the texts (or text types) included in the corpora. In other words, the sound advice of "knowing your data" becomes increasingly hard to follow when working with corpora that consist of billions of words from a huge number of different sources. It may therefore be argued that larger corpora have also brought about new kinds of problems (as noted, e.g., by Hiltunen, McVeigh & Säily 2017).

Finding unwanted items among search results is probably a very typical,almost everyday experience for many corpus linguists. Corpus compilers spend considerable effort regarding corpus annotation, markup, boilerplate removal, identification of duplicates or OCR errors. Similarly, scholars using corpora use increasingly refined methods when constructing elaborate query strings. Yet, achieving perfect precision and/or recall is still highly unlikely. We often need to exclude different kinds of search hits from further analyses, and the reasons for weeding out unwanted items canbe varied. Occasionally the occurrence of false positives is mentioned in research articles, perhaps in a footnote, but it may also be the case that much of the clean-up of irrelevant items is done silently.

For example, finding a search term within a quotation in a corpus text might justifiably give rise to exclusion of a token from further analysis. In fact, quotations can constitute a significant part of many corpora even in terms of their word count, yet their role overall in corpora has received little attention (see e.g. Rissanen 1992; Kaunisto 2017). Another related issue concerns the inclusion of various levels of linguistic annotation in corpora, which are often accepted as given especially by less experienced corpus linguists, but which may at times be less than helpful (see, e.g., Sinclair 2004; Archer 2012). Furthermore, the dispersion of tokens across the corpora can be a significant factor when assessing search results (see e.g. Gries 2008).

There are undoubtedly many types of persistent problems and messiness in corpus data that seasoned scholars have encountered and know about, but which are seldom specifically addressed. Yet beginning corpus users might benefit from learning about what may be regarded as tacit knowledge in corpus linguistics,and even the more advanced scholar may encounter issues new to them that have been addressed earlier. This workshop intends to tap into this knowledge by inviting papers on the following topics:

- false positives found in corpora; how to find them or assess their frequency in a corpus?
- the significance of identifying different types of unwanted items; how to deal with them and what are the risks if they are not identified?
- problems associated with categories built into corpus design and various types of linguistic annotation in corpora; to what extent can these seemingly helpful features encourage uncritical thinking or guide corpus users research?

It deserves to be mentioned that problematic aspects may be detected in individual corpora, and observing such infelicities as well as dealing with them is without question necessary and useful as the aim of such observations is to advance corpus linguistic endeavours on the whole. However, instead of focusing on corpus-specific issues, this workshop welcomes papers that reflect on general issues or their own experiences of, and mistakes in, corpus compiling and corpus-based research. In the collegial spirit of ICAME, this workshop is not intended as a forum for highlighting mistakes or shortcomings in fellow scholars' work.

References

Archer, Dawn. 2012. Corpus annotation: a welcome additionor an interpretation too far?, in Outposts of Historical Corpus Linguistics: From the Helsinki Corpus to a Proliferation of Resources, edited by Jukka Tyrkkö, Matti Kilpiö, Terttu Nevalainen & MattiRissanen. Studies in Variation, Contacts and Change in English 10. Helsinki: eVarieng. Available online at http://www.helsinki.fi/varieng/series/volumes/10/ archer/

Gries, Stefan Th. 2008. "Dispersions and adjusted frequencies in corpora". International Journal of Corpus Linguistics, 13:4, 403–437.

Hiltunen, Turo, Joe McVeigh & Tanja Säily. 2017. "How to turn linguistic data into evidence?", in Big and Rich Data in English Corpus Linguistics: Methods and Explorations, edited by Turo Hiltunen, Joe McVeigh &Tanja Säily. Studies in Variation, Contacts and Change in English 19. Helsinki: eVarieng. Available online at http://www.helsinki.fi/varieng/series/volumes/19/ introduction.html.

Kaunisto, Mark. 2017. "Multilingualism and quotations from a corpus-linguistic perspective: a case study of Samuel Taylor Coleridge's Biographia Literaria", in Challenging the Myth of Monolingual Corpora, edited by Arja Nurmi, Tanja Rütten, and Päivi Pahta, 220-238. Leiden: Brill.

Rissanen, Matti. 1989. "Three problems connected with the use of diachronic corpora". ICAME Journal 13: 16-19.Rissanen, Matti. 1992. "The diachronic corpus as a window to the history of English", in Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991, edited by Jan Svartvik, 185-205. Berlin and New York: Mouton de Gruyter.

Sinclair, John. 2004. Trust the Text: Language, Corpus and Discourse. London: Routledge.

Callies, Marcus (Universität Bremen)

Challenges in the annotation and analysis of learner corpora

In this talk I will highlight and discuss the special characteristics of learner corpus data and the challenges these may present for corpus compilation, annotation and analysis. Because learner corpus and SLA researchers use the data to study L2 production and development it is of utmost important that the data are valid, i.e. they represent "authentic" L2 production.

Texts contained in learner corpora have, by definition, been produced by bi- or even multilingual individuals, thus multilingual practices and phenomena induced by language contact, such as code-switching, foreignizing or calquing, are commonplace. These present challenges for annotation and analysis alike (Callies & Wiemeyer 2017). Learner corpora, especially those of academic texts, contain expert terminology, metalinguistic language use, e.g. examples ("mentioned items"), citations, and sometimes even whole abstracts or thematic summaries from other languages. Such instances do not represent 'genuine' learner production as they are typically taken over or copied from secondary sources. They can thus be considered unwanted items or "false positives" as their inclusion in word counts and concordance analyses will distort the data. They should thus be specifically tagged so that they can be excluded from analysis and frequency counts (Callies & Wiemeyer 2017: 90).

A further challenge that compilers and users of learner corpora have to deal with is unwanted lexical bias. This is introduced either by the topic of the task, or because learners use words or phrases from the task description, the writing prompt or other input material (see e.g. O'Donnell et al. (2013) for a description of how this may affect the use of lexical bundles in argumentative writing). It is important that researchers control for such effects because lexical variation, sophistication and complexity are often considered as proxies for L2 proficiency. Identifying lexical bias can be challenging, but if it is not discovered, its effects threaten the validity of the research findings. Words identified to cause lexical bias are either treated as stopwords, or L2 structures that are likely to have been brought about by lexical bias are excluded from the analysis. Similarly, task- and prompt-material may trigger the recurrent use of a whole grammatical construction. For instance, Callies (2008) notes an effect of the writing prompt on the occurrence of raising constructions, and Alexopoulou et al. (2015) discuss various task-effects on the use of relative clauses.

Finally, I will discuss the potential bias of certain annotation methods used in Learner Corpus Research (LCR), but also in other disciplines. LCR has been influenced by the "discourse of deficit" (Ortega 2013) that is still prevalent in much SLA research which is linked to the use of a monolingual native-speaker norm as the benchmark for the assessment of learner data. Error annotation thus often tends to be overly prescriptive. Creative and innovative but "non-standard" interlanguage forms (which are often contact-induced or formed on the basis of semantic or structural analogy to L2 input) may be considered "unwanted items" from an exonormative point of view, but they actually present valuable and highly interesting data for research into SLA and nativization processes in World Englishes (see e.g. Callies in press). In the ICE corpora family, such cases are described in the tagging manual for written texts in a section on "Normalizing the text" (see Nelson 2002).

References

Alexopoulou, T., Geertzen, J., Korhonen, A. & Meurers, D. 2015. Exploring big educational learner corpora for SLA research: Perspectives on relative clauses. *International Journal of Learner Corpus Research* 1(1), 96–129.

Callies, M. 2008. Easy to understand but difficult to use? Raising constructions and information packaging in the advanced learner variety. In G. Gilquin, M.B. Diez-Bedmar & S. Papp (eds.), *Linking Contrastive and Learner Corpus Research* (Language and Computers. Studies in Practical Linguistics, Band 66). Amsterdam: Rodopi, 201–226.

Callies, M. in press. Errors and innovations in L2 varieties of English: Towards resolving a contradictory practice. In G. Febel, K. Knopf, C. Nolte & M. Nonhoff (eds.), *Contradiction Studies: Mapping the Field*. New York: Springer.

Callies, M. & Wiemeyer, L. 2017. Multilingual speakers, multilingual texts: Multilingual practices in learner corpora. In A. Nurmi, T. Rütten & P. Pahta (eds.), *Challenging the Myth of Monolingual Corpora*. Amsterdam: Brill, 80–94.

Nelson, G. 2002. *International Corpus of English. Markup Manual for Written Texts*. https://www.ice-corpora.uzh.ch/dam/jcr:df7b1e8f-005f-4346-903b-c77b6c1da66a/written.pdf

O'Donnell, M. B., Römer, U. & Ellis, N. C. 2013. The development of formulaic language in first and second language writing: Investigating effects of frequency, association, and native norm. *International Journal of Corpus Linguistics* 18, 83–108.

Ortega, L. 2013. SLA for the 21st century: Disciplinary progress, transdisciplinary relevance, and the bi/multilingual turn. *Language Learning* 63 (Suppl. 1), 1–24.

Egbert, Jesse (Northern Arizona University), Tove Larsson (Uppsala University), and Douglas Biber (Northern Arizona University)

Research design in quantitative corpus linguistics: Critical reflections and suggested improvements

This paper focuses on research design in quantitative corpus linguistics. The impetus for this project is our observation that there has been a general trend in corpus linguistics away from a focus on actual language data. This observation is corroborated by preliminary findings from a survey of articles published in major corpus linguistics journals in 2009 and 2019. Results from that study reveal that during the past decade there have been statistically significant increases in the proportion of the prose in the results and discussion section devoted to statistical reporting (p = .02, d = 1.19) and in the number of distinct statistical techniques used (p = .02, d = 1.22). The average quantitative corpus study relies on more actual language data than research in other areas of linguistics. Despite this, we propose that corpus linguists are increasingly focused on quantifying linguistic patterns at the expense of linguistic analysis. This trend seems to be accelerated by a number of factors that are conspiring to draw quantitative corpus linguists away from analysis of actual language data. These factors emerge in every major step of the research process. Major factors that distract from the goal of linguistic analysis in quantitative corpus research include increases in corpus size, misalignment of research questions and methods, observational units with limited linguistic validity, uninformative language variables, overly complex statistical models, and lack of qualitative interpretation. We believe there is compelling evidence that these factors have a tendency to create layers of distance between the researcher and the language data in the corpus under study.

We will begin the presentation by introducing and illustrating the trend away from linguistic analysis in quantitative corpus linguistics and make a case for corpus linguistics to return to the primary goal of language description. We will then turn to a detailed description of two major research design decision points that seem to be drawing corpus linguists away from descriptive linguistic analysis: the use of statistical models and qualitative analysis and interpretation of language data. We propose that sophisticated statistical models can, at times, create unnecessary distance between the researcher and the data, especially when the statistical analysis is not coupled with qualitative linguistic analysis. We will propose a minimally sufficient approach to statistical analysis in which the researcher uses statistics that are no more nor less sophisticated than necessary to answer the research questions. We will then stress the importance of returning to the actual language to interpret and explain quantitative patterns and fully explain them to the reader. We will demonstrate good examples of this in the literature and demonstrate the potential perils of drawing conclusions from quantitative data without conferring with the actual language in texts themselves. We will conclude the paper by introducing a series of suggestions for improvements to research design in quantitative corpus linguistics that will aid researchers in their attempts to make quantitative corpus-based research more language focused and linguistically informative.

Hartmann, Stefan (University of Bamberg)

Open Corpus Linguistics: Overcoming Rissanen's problems (and others) with open data

Corpus linguists are often faced with recurrent problems related to issues such as representativity, sample size, or the trade-off between precision and recall. In this paper, I argue that virtually all problems that corpus linguists face on a regular basis are somehow connected to transparency and openness (or the lack thereof). Take, for instance, the three problems outlined in Rissanen's (1989) seminal paper:

- "The philologist's dilemma" refers to the problem that corpus linguists tend to limit their view to key words in context, rather than taking full texts into account. In diachronic studies, this might even entail that researchers study phenomena in earlier stages of a language that they are not sufficiently familiar with. However, for many corpora, the full texts are not readily available. Instead, users can only obtain KWIC results via an online interface. In such cases, "the philologist's dilemma" is an unavoidable consequence of usage restrictions.
- "God's truth fallacy" pertains to the issue of representativeness: A researcher may be led to believe that the corpus they use accurately represents the linguistic variety that it is intended to represent. Rissanen (1989) therefore suggests to keep corpora open-ended. Arguably the best way of doing so is by making corpora openly available, which allows researchers to add their own annotations or modify existing ones.
- "The mystery of vanishing reliability" refers to the problem that very fine-grained annotation schemes may be problematic because any increase in the number of parameter values leads to a decrease in the number of tokens associated with each parameter value, as long as the size of the corpus is kept constant. However, this is not necessarily a problem: Fine-grained annotations can easily be broken down to more coarse-grained ones (but not vice versa). But this is only possible if the data and annotations are readily available.

All three problems cannot be directly solved: The first two are problems of corpus linguists, rather than of corpus linguistics, and the third is an empirical phenomenon that is simply unavoidable. All three, however, call for keeping the limits of corpus linguistics in mind, and open corpus data can help in doing so: It encourages researchers to work with and enrich the data, rather than seeing the existing corpus as an authoritative resource, thus avoiding the first two problems and providing good ways of dealing with the third. Thus, in line with the recent trend towards openly sharing research data (Zwaan 2017, Berez-Kroeker et al. 2018), I argue that corpus linguists should a) pursue the ideal of making the full annotated texts of the corpora they compile available to the public free of charge under permissive licenses and b) keep the aspect of transparency and openness in mind when choosing the empirical basis of their analyses, preferring open resources like the Open American National Corpus (OANC) over more widely-used, but less transparent resources whenever possible.

References

Berez-Kroeker, Andrea L., Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, David I. Beaver, Shobhana Chelliah, Stanley Dubinsky, Richard P. Meier, Nick Thieberger, Keren Rice & Anthony C. Woodbury. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 56(1). 1–18.

Rissanen, Matti. 1989. Three problems connected with the use of diachronic corpora. *ICAME Journal* 13. 16–19.

Zwaan, Rolf A., Alexander Etz, Richard E. Lucas & M. Brent Donnellann. 2017. Making replication mainstream. *Behavioral and Brain Sciences*. doi:10.1017/S0140525X17001972.

Hiltunen, Turo (University of Helsinki)

Issues in using *British Library Newspapers* as a corpus

The availability of massive text archives holds great promise for corpus linguis- tic work, but at the same time they also present considerable methdological chal- lenges for users (see e.g. Hiltunen et al. 2017). This paper discusses some specific challenges encountered when doing corpus linguistic research using the *British Library Newspapers* database,[1] which contains several million pages from national and regional newspapers from Britain between the 18th and the 20th centuries, and considers some potential solutions for them. Of particular interest in the present study are issues relating to the study of linguistic variation and register analysis.

Previous studies have already identified issues in the use of this database for research in history, cultural studies and Digital Humanities, including the consid- erable amount of OCR (Optical Character Recognition) errors (e.g. Gregory et al. 2016); these errors may cause problems in precision and recall, and it is not ob- vious what the best way of dealing with them would be. Another major issue is the representativeness and balance of data from text archives: even when using the full-text data instead of the native web interface, *BL Newspapers* clearly lacks the tidiness and balance of smaller corpora like ARCHER (Hundt and Leech 2012), which would greatly facilitate the interpretation of quantitative results (see e.g. Davies 2019). Another issue related to balance is the fact that content is missing from a number of years (Nicholson 2012), which needs to be taken into account in diachronic analyses. Finally, despite the fact that *BL Newspapers* contains metadata about different text categories (e.g. *news*, *editorials*, and *sports*), these categories are not necessarily directly suitable for comparing differences in language use at the level of register and sub-register (Biber and Gray 2013). This paper illustrates these and related issues, considers to what extent the may have an impact on the quan- titative results, and discusses and evaluates some ways of dealing with them.

References

Biber, Douglas and Bethany Gray (2013). “Being Specific about Historical Change: The Influence of Sub-Register”. In: *Journal of English Linguistics* 41.2, 104–134.

Davies, Mark (2019). “Corpus-based studies of lexical and semantic variation: The importance of both corpus size and corpus design”. In: *From Data to Evidence in English Language Research*. Ed. by Terttu Nevalainen Carla Suhr and Irma Taav- itsainen. Leiden: Brill, 66–87.

Gregory, Ian Norman, Paul David Atkinson, Andrew Hardie, Amelia Joulain-Jay, Daniel Kershaw, Catherine Porter, Paul Edward Rayson, and Christopher John Rupp (2016). “From digital resources to historical scholarship with the British Library 19th Century Newspaper Collection”. In: *Journal of Siberian Federal Uni- versity: Humanities and social sciences* 9.4, 994–1006.

Hiltunen, Turo, Joe McVeigh, and Tanja Säily (2017). “How to turn linguistic data into evidence?” In: *Big and Rich Data in English Corpus Linguistics: Methods and Explorations*. Ed. by Turo Hiltunen, Joe McVeigh, and Tanja Säily. Studies in Vari- ation, Contacts and Change in English. Helsinki: Research Unit for Variation, Contacts, and Change in English.

Hundt, Marianne and Geoffrey Leech (2012). ““Small is beautiful”: On the value of standard reference corpora for observing recent grammatical change”. In: *The Oxford Handbook of the History of English*. Ed. by Elizabeth Traugott and Terttu Nevalainen. Oxford: Oxford University Press, 175–188.

Nicholson, Bob (June 2012). “Counting Culture; or, How to Read Victorian News- papers from a Distance”. In: *Journal of Victorian Culture* 17.2, 238–246.

[1] See https://www.gale.com/intl/c/british-library-newspapers-part-i.

Kaunisto, Mark (Tampere University)

Proper names as potentially problematic items in corpora

There are many types of elements found in corpora which, while they are perfectly representative of normal language use, may turn present themselves as problematic when investigating patterns of language use. For example, Rissanen (1992) paid attention to the role of quotations in historical sermons, and observed that biblical quotations contained relatively higher proportions of pronouns compared to the rest of the sermon texts. In similar vein, Kaunisto (2017) examined the overall number of words from quotations as well as foreign language passages in Coleridge's *Biographia Literaria,* and found that approximately 12 per cent of the entire word count of the book was made of items which do not exactly represent Coleridge's own choices of linguistic patterns; in many instances, the quotations even represented language written hundreds of years before Coleridge's time. In other words, texts may sometimes feature items which have the potential of skewing the results. It is therefore of interest to examine more closely the characteristics of such items and to assess seriousness of such effects on different types of linguistic analyses.

The occurrence of proper names may arguably be identified as an issue that may cause some difficulty in some types of corpus analyses. Even though corpus users are probably well aware that items found in proper names – e.g. words found in titles of books, articles, films, etc. – may be regarded as frozen items and would normally be excluded from further analysis because the choice of words in such instances was usually made by someone else than the authors themselves. However, in automated collocational analyses it may not be possible to make use of part-of-speech tags to separate between instances of words where the word has been a part of a proper name or not. The situation is somewhat easier if we need to distinguish between proper nouns and common nouns, inasmuch as we can rely on the taggings in this regard. Achieving accuracy in tagging these items has But as regards, for example, the occurrences of adjectives within proper names – e.g. *magical* in *Magical Mystery Tour* – no annotation is necessarily available for us to separate such instances from regular uses of the adjective, and close manual inspection is therefore needed to exclude such items from the analysis.

In this paper, I will examine through different types of examples how the frequencies of proper name uses may play an important role in studies of, for example, near-synonyms, and how the occurrence of proper name use may show different degrees of prominence of words in different registers. Overall, the main argument is that due caution must be given to such items and sufficient manual inspection of concordance lines is needed to avoid the possibility of misinterpreting corpus data.

References

Kaunisto, Mark. 2017. "Multilingualism and quotations from a corpus-linguistic perspective: a case study of Samuel Taylor Coleridge's *Biographia Literaria*", in *Challenging the Myth of Monolingual Corpora*, edited by Arja Nurmi, Tanja Rütten, and Päivi Pahta, 220-238. Leiden: Brill.

Rissanen, Matti. 1992. "The diachronic corpus as a window to the history of English", in *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991*, edited by Jan Svartvik, 185-205. Berlin and New York: Mouton de Gruyter.

Miletic, Filip, Anne Przewozny-Desriaux and Ludovic Tanguy

CLLE, CNRS & University of Toulouse, Toulouse, France

{filip.miletic, anne.przewozny, ludovic.tanguy}@univ-tlse2.fr

Modeling fine-grained sociolinguistic variation: the promises and pitfalls of Twitter corpora and neural word embeddings

Recent work in natural language processing suggests that the vast amount of data in social media can provide significant insights into language variation (Eisenstein, 2018), and that sophisticated statistical models can detect phenomena such as semantic change on a scale which defies manual analysis (Tahmasebi et al., 2018). However, although these approaches are promising, most of them are yet to provide substantive descriptive contributions (Boleda, 2020).

We examine the usability of these methods in corpus-based sociolinguistic research by investigating contact-induced semantic shifts in Quebec English, i.e., English words used with meanings typical of phonologically similar French words. Our aim is to expand on anecdotal evidence (Fee, 2008; Boberg, 2012; Rouaud, 2019) and explore underlying quantitative usage patterns. We analyze 45 previously described examples in a custom-built, 1.3-billion-word corpus of English tweets posted by users from Montreal, Toronto and Vancouver (Miletic et al., 2020). For each analyzed word, we use BERT (Devlin et al., 2019), a neural network trained on a large generic corpus of English, to produce vector representations of individual occurrences. These are then used to identify clusters of tweets sharing a similar meaning of the target word.

A detailed manual analysis shows that contact-related meanings – e.g. *deceive* 'disappoint' and *souvenir* 'memory' – typically appear in clusters where the majority of tweets was posted by users from Montreal, the only analyzed city with a large French-speaking population. Conversely, conventional meanings – here, *deceive* 'mislead' and *souvenir* 'memento' – mostly occur in clusters with a balanced regional distribution. An additional analysis extended to all languages shows that users in contact-related clusters produce more French tweets than those who use conventional meanings. Overall, contact-induced semantic shifts appear to involve a modified sense distribution (rather than a completely altered meaning) and represent a variation in usage related to bilingualism (rather than established regional variants).

However, we only arrived at these conclusions after controlling for a variety of false positives. Some clusters capture uses that are specific to Montreal, but unrelated to language contact. This includes proper nouns (*affirmations* referring to an album by a Montreal rapper) and cultural factors like the city's thriving IT industry (*library* denoting resources in software development). Other clusters involve French homographs rather than English semantic variants (*dossier* in *appel de dossiers* 'call for submissions'). This is due to borrowing and codeswitching, which are overall rare, but relatively more frequent in Montreal. Finally, some clusters reflect structural patterns (tweet- final tokens grouped together due to BERT's sensitivity to position) or idiolectal preferences.

Our results show that large corpora and statistical models can help advance the description of complex sociolinguistic phenomena. But even when considerable effort is put into data collection and filtering, corpus analyses can be skewed by topical and quantitative biases, preprocessing issues, and inherent limitations of computational models. It is vital to control for these factors before interpreting large-scale quantitative results. An ongoing sociolinguistic survey, based on a well-established variationist protocol (Przewozny et al., 2020), will help us further address these issues by comparing Twitter-based analyses to real-life sociolinguistic behaviors.

References

Boberg, Charles. (2012). English as a Minority Language in Quebec. *World Englishes* 31 (4): 493–502.

Boleda, Gemma. (2020). Distributional Semantics and Linguistic Theory. *Annual Review of Linguistics* 6 (1): 213-234.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT 2019*, 4171-4186.

Eisenstein, Jacob. (2018). Identifying Regional Dialects in On-Line Social Media. In Boberg, Charles, John Nerbonne, and Dominic Watt (eds.), *Handbook of Dialectology*, Hoboken, NJ: Wiley Blackwell, 368-383.

Fee, Margery. 2008. French Borrowing in Quebec English. *Anglistik: International Journal of English Studies* 19 (2): 173–88.

Miletic, Filip, Anne Przewozny-Desriaux, and Ludovic Tanguy. (2020). Collecting Tweets to Investigate Regional Variation in Canadian English. In *Proceedings of LREC 2020*, 6255–6264.

Przewozny, Anne, Cécile Viollain, and Sylvain Navarro. (2020). *The Corpus Phonology of English: Multifocal Analyses of Variation*. Edinburgh: Edinburgh University Press.

Rouaud, Julie. 2019. *Lexical and phonological integration of French loanwords into varieties of Canadian English since the seventeenth century*. PhD thesis, Université Toulouse – Jean Jaurès.

Tahmasebi, Nina, Lars Borin, and Adam Jatowt. (2018). Survey of Computational Approaches to Lexical Semantic Change. *Preprint at ArXiv 2018.*

Sundberg, Daniel (Linnaeus University)

Corpus Categories: What and for whom? When special corpora meet general corpora in comparative studies in literature

Fiction is inherently messy to work with. Not due to the material itself, but rather due to the field that surrounds it and the needs of different theoretical approaches to the reading of the materials. In the decades since Leech & Short's *Style in Fiction* (1981) the intersection between linguistics and literature has continued to develop. Mahlberg (2007) has made excellent descriptions of how the relationship between corpus linguistics and literary theory can be approached and the framework needed to successfully make use of our methodological resources within the field of literature. Biber (2011) provides interesting examples of different ways corpora have been used in the study of literature and shows how methods such as keyword analysis, n-grams, and collocations have been at the center stage. However, the use of large corpora for comparative studies within literature remains problematic, as these corpora were rarely constructed for this purpose.

Special corpora, defined by Tognini Bonelli (2012) as a corpus where the selection is not made to be representative of a language but of a specific use-case, play an important role here as literature corpora are often designed to be representative of an author, a period, or a genre. This makes the categorization procedures very different from the procedures used in general corpora. While categorization in large corpora may make use of broad text type categorizations in combination with temporal and spatial categorization, as in the BNC and the COHA/COCA, special corpora are often interested in other category tags. The temporal and spatial categorization of texts along with the text type becomes central for comparative studies of literature, for instance, a single American author of fiction active during the 1920s to the 1960s being contrasted with American fiction written during the 1920s to 1960s (as in Sundberg & Nilsson forthcoming).

As the interpretation of data begins, further questions regarding the categorizations arise, often to do with genre and style, and one must consider whether our American author active during the 1920s to 1960s was a modernist, if they wrote autobiographically, which genre conventions they adhered to and so forth. Comparing this author to any material matching the temporal and spatial categorization while tagged as "fiction" becomes problematic, especially when presenting to an audience who is mainly engaged with those other aspects of the author's work rather than the "when and where", for instance at a literary conference (Sundberg 2018, 2019). Categorization within corpora is a well-researched topic which has produced multiple excellent methods of approach, for instance using keywords (Özgür, Özgür & Güngör 2005), named-entity recognition (Sahin et. al 2017) or machine learning (Fabrizio 2002), but these are of limited use when the desired categorizations are based on features not directly tied to the language itself.

As this intersection becomes more popular, the need for a discussion on how these uses of corpora as contrastive, or comparative, resources beyond language variants and variation becomes important. How could this new arena influence our categorization habits, and what are the consequences of deeper categorization of fictional texts?

References

Biber, D. (2011). "Corpus linguistics and the study of literature: Back to the future?", The Scientific Study of Literature, 1, pp. 15-23.

Fabrizio, S. (2002). "Machine learning in automated text categorization", ACM Computer Survey 34, 1. Pp. 1–47. DOI: https://doi.org/10.1145/505282.505283.

Leech, G. and Short, M. (1981). *Style In Fiction*. London: Routledge Taylor & Francis Group.

Mahlberg, M. (2007). "Corpus Stylistics: Bridging the gap between linguistic and literary studies” in Hoey, M., 2007. *Text, Discourse And Corpora*. London: Continuum, pp. 219 – 246.

Sahin, H. Bahadir, C. Tirkaz, E. Yildiz, M. Tolga Eren, and O. Sonmez. (2017). ”Automatically annotated Turkish corpus for named entity recognition and text categorization using large-scale gazetteers”. Available at http://arxiv.org/abs/1702.02363.

Sundberg, D and Nilsson, J. (2021). “A Corpus Stylistic Analysis of Hemingway’s Literary Production” to be published in *The Hemingway Review*, Spring Issue 2021.

Sundberg, D. (2019). “’O Fudge, the Looks of the Girls’: A corpus-driven analysis of the female role in F. Scott Fitzgerald’s Fiction”. 15th F. Scott Fitzgerald Society Conference June 24-29 2019. Toulouse, France.

Sundberg, D. (2018). “’Cultivating One True Sentence’: A corpus stylistic analysis of Hemingway’s language”. XVIII Hemingway Society Conference July 22-28 2018. Paris, France.

Tognini Bonelli, E. (2012). “Theoretical overview of the evolution of corpus linguistics” in O'Keeffe, A. and McCarthy, M., 2012. *The Routledge Handbook Of Corpus Linguistics*. Milton Park, Abingdon, Oxon: Routledge, pp. 14 – 28.

Özgür A., Özgür L, and Güngör T. (2005). ”Text Categorization with Class-Based and Corpus-Based Keyword Selection”. In: Yolum ., Güngör T., Gürgen F., Özturan C. (eds) Computer and Information Sciences - ISCIS 2005. ISCIS 2005. Lecture Notes in Computer Science, vol 3733. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11569596_63

Tyrkkö, Jukka (Linnaeus University) and Sophie Raineri (Paris Nanterre University)

Empirical perspectives on the reliability and accuracy of collaborative pragmatic annotation

Corpora can be annotated with a wide variety of different data, ranging from shallow morphosyntactic tagging to deep syntactical parsing, and from synonyms and semantic categories to named entities and reference chain identification. The general usefulness of linguistic annotation is rarely a topic of much controversy, but the methods used and the reliability thereof continue to raise questions. While some types of analysis can be performed algorithmically with increasing accuracy, such as part-of-speech tagging, others continue to require the efforts of human annotators (see Hovy & Lavid 2010). The latter types of annotation tasks raise particular concerns, because while erroneous annotations do occur with computational methods, the errors are usually systematic and thus easy to account for and correct (see Archer 2012). By contrast, human annotators tend to work more inconsistently, and their performance is believed to rely extensively on how they were instructed and how well they understood the instructions, as well as on random contextual and situational variables, which may lead to significant unreliability even at the within-annotator level (see, e.g., Larsson et al 2020). In the corpus linguistic setting, an added challenge comes from the sizes of modern corpora, which would typically make it impossible for a single human annotator to work through the entire corpus and therefore requires the involvement of multiple annotators.

Out of all the different types of linguistic annotation, pragmatic annotation may be the most prone to annotator errors and inconsistencies (see Archer et al. 2008). This is particularly true of pragmatic features such as functional units within spoken or written samples, and most especially if the identification of such units is based on human competence of understanding language in culture-dependent and/or intertextual contexts, such as humour or different types of storytelling (see, e.g., Alsop 2016 and Alsop et al. 2013). At the same time, annotations of this type are potentially very useful, as they would allow researchers to focus on specific sections of the corpus, or to disregard specific sections, without first having to manually analyse the corpus (see, e.g., Maynard & Leichter 2007). What, then, can we realistically expect from collaborative annotation projects in terms of reliability and accuracy? How much training is needed and does training substantially improve the results?

We tested the reliability and accuracy of collaborative pragmatic annotation with a total of 18 graduate students from Linnaeus University and 35 from Paris Nanterre University. Using a controlled selection of 12 subsets of extracts from 91 speeches of the *Diachronic Corpus of Political Speeches*, the students were asked to annotate three types of pragmatic segments, namely OPENINGS, NARRATIVES, and HUMOUR. In the first phase of the study, 43 students annotated segments with minimal guidance, while in the second phase, a new group of 10 students received extensive instruction and engaged in collaborative discussion prior to the annotation task. Each subset of data was annotated by a total of 8 students: 5 in phase one and 3 in phase two. The research design allowed the analysis of the effects of guidance on consistency and reliability of the annotation task, in relation to the individual annotators and the different types of annotation tasks (see Banerjee et al 1999, Potter & Levine-Donnerstein 1999).

Our paper will discuss the research design and the two different annotator training models, present and compare the results of the two phases of annotation, discuss relevant statistical aspects of inter-annotator reliability assessment (with special reference to Bowker's test for consistency and Krippendorff's alpha for agreement), and conclude with some recommendations for future collaborative annotation projects.

References

Alsop, Siân. 2016. The 'humour' element in engineering lectures across cultures: an approach to pragmatic annotation. In Maria Jose Lopez-Couso, Belen Mendez-Naya, Paloma Nunez-Pertejo & Ignacio M Palacios-Martinez (eds.) *Corpus Linguistics on the Move. Exploring and Understanding English through Corpora*. London: Brill. 337–361.

Alsop, Siân, Emma Moreton & Hilary Nesi. 2013. The uses of storytelling in university engineering lectures. *ESP Across Cultures*, 10. 7–19.

Archer, Dawn, Jonathan Culpeper & Matthew Davies. 2008. Pragmatic annotation. In Kytö, Merja & Anke Lüdeling (eds.) *Corpus Linguistics: An International Handbook*. Mouton de Gruyter, 613-642.

Archer, Dawn. 2012. Corpus annotation: A welcome addition or an interpretation too far? In Jukka Tyrkkö, Matti Kilpiö, Terttu Nevalainen & Matti Rissanen (eds.) *Outposts of Historical Corpus Linguistics: From the Helsinki Corpus to a Proliferation of Resources*. (Studies in Variation, Contacts and Change in English 10). Helsinki: Varieng. Available online at <https://varieng.helsinki.fi/series/volumes/10/archer/>

Banerjee, Mousumi, Michelle Capozzoli, Laura McSweeney & Debajyoti Sinha. 1999. Beyond kappa: A review of interrater agreement measures. *The Canadian Journal of Statistics*, 27(1), 3-23.

Diachronic Corpus of Political Speeches (DCPS). In progress, expected 2022. Corpus compiled by Jukka Tyrkkö, Sophie Raineri & Jenni Riihimäki at Linnaeus University, Paris Nanterre University, and Tampere University. Freely available under Creative Commons license BY-NC-ND.

Hovy, Eduard & Julia Lavid. 2010. Towards a 'science' of corpus annotation: A new methodological challenge for corpus linguistics. *International Journal of Translation*, 22:1. 13–36.

Larsson, Tove, Magali Paquot and Luke Plonsky. 2020. Inter-rater reliability in learner corpus research. *International Journal of Learner Corpus Research*, 6(2). 237–251.

Maynard, Carson and Sheryl Leichter. 2007. Pragmatic annotation of an academic spoken corpus for pedagogical purposes. In Fitzpatrick, Eileen (ed.) *Corpus Linguistics Beyond the Word: Corpus Research from Phrase to Discourse*. Amstredam & New York: Rodopi. 107–116.

Potter, W. James & Deborah Levine-Donnerstein. 1999. Rethinking validity and reliability in content analysis. *Journal of Applied Communication Research*, 27(3), 258–284.

Vetter, Fabian (University of Bamberg)

Register variation & comparability in parallel corpora

Ever since the beginnings of English corpus linguistics, the distribution of a linguistic variable is often compared in multiple corpora and/or parts of such. While the outer boundaries of corpora are usually defined geographically, internally these are typically structured into registers. This latter practice is certainly justified, as it has been shown that register has an effect on a great number of linguistic phenomena.

What is problematic, however, is that the lines along which corpus texts are divided into registers are often blurry and the commonly encountered sparsity of metadata does not allow for alternative categorizations. Instead of treating register as a multivariate phenomenon, so that the grouping of texts can be adjusted according to specific study goals – as suggested by Leitner (1992) and Sigley (1997) – linguists usually have little choice but to rely on predefined text category labels. As a result, the effects certain text external factors might have on a linguistic variable cannot be explored. Another issue resulting from this practice concerns the use of comparable or parallel corpus families (i.e. corpora that share the same sampling scheme) such as the BROWN corpus family or the International Corpus of English (ICE). Identical register labels in different corpora subtly suggest that these are unconditionally comparable. In the absence of other obvious distorting factors (e.g. diachronic variation, see also Hundt 2015 for a discussion on distorting factors in ICE), attributing observed differences in the distribution of a linguistic variable to regional variation then stands to reason.

This study mainly addresses the latter issue, illustrates that some registers in ICE, despite identical sampling schemes and register labels, contain different text types and argues that these differences decrease the comparability. To this end, a blend of quantitative and qualitative methods is used to detect and analyse instances where differences in sampling occur.

Similar to other studies that have shown that vector representations of texts can be used to reliably cluster texts by register (cf. Gries et al. 2011, Fang & Cao 2015, Tang & Cao 2015), this study utilizes metric Multidimensional Scaling to detect such differences. The vector representations in this study are based on the frequencies of parts-of-speech monograms. Where distinct patterns for some registers emerge, the results are investigated qualitatively through close reading, and reviewing metadata and, where procurable, facsimiles of the original texts.

It is shown that some registers in some components of ICE exhibit undocumented differences in terms of sampled text types. As such differences would otherwise remain undetected, it is consequently argued that corpus texts should be annotated with text external criteria (e.g. following Biber & Conrad 2009). While these compositional disparities are practically unavoidable, especially when data from various regional varieties of English are collected, such an annotation would render them at least visible and they could be taken into account in corpus comparisons.

References

Biber, Douglas & Susan Conrad. 2009. *Register, genre, and style* (Cambridge textbooks in linguistics), 1st edn. Cambridge: Cambridge Univ. Press.

Fang, Alex C. & Jing Cao. 2015. *Text Genres and Registers: The Computation of Linguistic Features*. Berlin, Heidelberg: Springer.

Gries, Stefan T., John Newman & Cyrus Shaoul. 2011. N-grams and the clustering of registers. *Empirical Language Research* 5(1).

Hundt, Marianne. 2015. World Englishes. In Douglas Biber & Randi Reppen (eds.), *The Cambridge Handbook of English Corpus Linguistics*, 381–400. Cambridge: Cambridge University Press.

Leitner, Gerhard. 1992. International Corpus of English: Corpus design - problems and suggested solutions. In Gerhard Leitner (ed.), *New directions in English language corpora*: *Methodology, results, software developments* (Topics in English linguistics 9), 33–64. Berlin, New York: Mouton de Gruyter.

Sigley, Robert. 1997. Text Categories and Where You Can Stick Them: A Crude Formality Index. *International Journal of Corpus Linguistics* 2(2). 199–237.

Tang, Xiaoyan & Jing Cao. 2015. Automatic Genre Classification via N-grams of Part-of-Speech Tags. *Procedia - Social and Behavioral Sciences* 198. 474–478

Workshop 2: Crossing Language and Discipline Boundaries

Convenors: Anna Čermáková (University of Cambridge, UK / Charles University, Czech Republic), Hilde Hasselgård (University of Oslo, Norway), Markéta Malá (Charles University, Czech Republic), Denisa Šebestová (Charles University, Czech Republic)

We would like to propose a contrastive pre-conference workshop at ICAME 42. Contrastive corpus-based linguistics is closely tied to the history of ICAME and as such has traditionally presented a distinctive perspective at ICAME conferences. The first contrastive pre-conference workshop in the series was organised ten years ago in 2011 at ICAME 32 in Oslo by Karin Aijmer and Bengt Altenberg (Aijmer & Altenberg 2013). But already in 1993 at ICAME in Zürich Stig Johansson "presented his vision of a corpus-based project that was to begin a new era in contrastive linguistics and translation studies" (quoted from the call for the 2011 workshop). To honour this important anniversary, we have invited Prof. Karin Aijmer to open our workshop and reflect on those ten years.

In 2020, at the experimental digital format of ICAME in Heidelberg, the contrastive workshop was one of the two that went ahead despite the circumstances – confirming the strength of the community. These workshops have become an immensely important forum for the exploration of both theoretical and practical issues in corpus-based contrastive linguistics. Each workshop has been followed by a publication, the latest being the forthcoming in 2021 *Time in Languages, Languages in Time* (ed. by A. Čermáková, H. Hasselgård, T. Egan & S. Rørvik) based on the Neuchâtel workshop (ICAME40), preceded by *Corpora et Comparatio Linguarum: Textual and Contextual Perspectives* (ed. by S.O. Ebeling & H. Hasselgård, 2018) based on the ICAME 38 workshop in Prague – see the appendix for a full list.

Through the lens of language comparison, we enrich our understanding of the complexities of language in general and the languages compared in particular. Contrastive corpus-based studies have become an established research strand within the field of English corpus linguistics, and the proposed ICAME workshop, like all its predecessors, invites contributions in which English is compared to at least one other language. As corpus research itself is moving beyond the space it has traditionally occupied, contrastive studies need to reflect this development. We would like to echo the overall conference theme of crossing boundaries, since contrastive studies by definition cross the boundaries between languages. Furthermore, we would particularly welcome contributions that cross boundaries between traditional linguistic disciplines as well as between linguistics and other disciplines.

References

Previous contrastive pre-conference workshops at ICAME have produced the following publications:

Čermáková, A., Egan, T., Hasselgård, H. & Rørvik, S. (eds) (2021). *Time in Languages, Languages in Time.* Amsterdam: John Benjamins. (ICAME 40, Neuchâtel 2019)

Ebeling S.O. & Hasselgård, H. (eds) (2018). *Corpora et Comparatio Linguarum: Textual and Contextual Perspectives.* Bergen Language and Linguistics Studies (BeLLS) Vol 9(1). (ICAME 38, Prague 2017)

Janebová, M., Lapshinova-Koltunski, E. & Martinková, M. (eds) (2017). *Contrasting English and other Languages through Corpora*. Newcastle: Cambridge Scholars Publishing. (ICAME 37, Hong Kong 2016)

Egan, T. & Dirdal, H. (eds) (2017). *Cross-linguistic Correspondences.* Amsterdam: John Benjamins. (ICAME 36, Trier 2015)

Ebeling S.O. & Hasselgård, H. (eds) (2015). *Cross-linguistic Perspectives on Verb Constructions*. Newcastle: Cambridge Scholars Publishing. (ICAME 35, Nottingham 2014)

Aijmer, K. & Hasselgård, H. (eds) (2015). *Cross-linguistic Studies at the Interface Between Lexis and Grammar*. Special issue of Nordic Journal of English Studies (Volume 15:1). (ICAME 34, Santiago de Compostela 2013)

Altenberg, B. & Aijmer, K. (eds) (2013). *Text-based Contrastive Linguistics*. Special issue of *Languages in Contrast* (Vol. 13:2). (ICAME 33, Leuven 2012)

Aijmer, K. & B. Altenberg (eds) (2013). *Advances in Corpus-based Contrastive Linguistics. Studies in honour of Stig Johansson*. Amsterdam: John Benjamins. (ICAME 32, Oslo 2011)

Aijmer, Karin

The contrastive workshop - the first ten years

The first contrastive workshop was organized in Oslo 2011 in honour of the late Professor Stig Johansson who was an enthusiastic and inspiring pioneer in the field corpus-based contrastive linguistics.It has since been followed by contrastive workshops at ICAME -conferences, for example, in Leuven, Santiago de Compostela, Nottingham, Hong Kong, Prague and Neuchâtel. During a 10-year-period we have witnessed a dramatic development of corpus-based contrastive linguistics to more languages and language pairs studied in a contrastive perspective, as well as the creation of new corpora and corpus tools. As is shown by the contributions to the workshops, linguistic expressions are now studied on many different levels including pragmatics, text, discourse and phraseology. The presentation will give a bird's eye view of the development of the field by looking back at the preceding workshops. I will end with some reflections on the future of corpus-based contrastive linguistics.

Bourgoin, Charlotte, Kristin Davidse, Karen Lahousse (KU Leuven)

A framework for cross-linguistic study of the information structure of English it- and French c'est-clefts

In English and French, clefts are a productive resource serving information structural functions (Filppula 2009), for which different typologies have been drawn up in the respective literatures. These seek to identify the different information statuses the complement of the matrix, the 'clefted element', e.g. *he* in (1), and the cleft relative clause (CRC), *who invited me*, may have.

(1) A: did you meet Fuller?
B: y/es# **it was he who inv\ited me**# (LLC-1)

The existing typologies tend to conflate two layers of information structure, viz. 'new' versus 'given' information and 'accented' versus 'non-accented' information. This conflation characterizes the two types posited in Prince's (1978) influential typology. The *stressed-focus* cleft is defined as having a new clefted element with strong stress and a CRC with given and weakly stressed information. The *informative-presupposition* cleft is said to have an unstressed, given clefted element and a CRC with new and normally stressed information. Corpus-based studies in both English (Delin 1990, Collins 2006) and French (Doetjes et al. 2004, Avanzi 2011, Mertens 2012) have revealed greater prosodic variation which is not in fact bound to Prince's two types of cleft, but they continue to work with a modified form of Prince's binary typology.

In this paper, we argue that a model of the information structure of clefts has to make a principled distinction between discourse-familiarity (Kaltenböck 2005) and focal versus non-focal status as marked by prosody (Halliday & Greaves 2008). Discourse-familiarity addresses the question of whether the referents of the clefted element and the CRC are directly evoked in or inferable from the preceding discourse, i.e. discourse-given or new. Focus-assignment involves the speaker's marking by tonic prominence of the most salient information within prosodically delineated information units. With this fine-grained analytical framework we study 500 tokens of clefts extracted from audio files of the London-Lund Corpus (LLC-1) and the Corpus de Référence du Français Parlé (CRFP). For the prosodic analysis, we combine instrumental and auditory analysis to inventory the *number* and *location* of nuclei coding information focus.

Our analysis of the English and French data has revealed that in terms of *discourse-givenness*, four types have to be distinguished, new – given, given – new, new – new, and given – given. The last type, illustrated in (2), had hitherto been overlooked.

(2) A: from Marlborough she has hit Reading at half past eight in the morning
[…]
A: it is about eight th\irty# that my mother has gotten st\uck# in traffic trying to get into R\eading# (LLC-1)

With each option of this quaternary typology, we correlate the unmarked, most frequent pattern of focus assignment as well as the marked, less common patterns. For instance, an anaphoric pronoun as clefted element is typically non-focal, e.g. (1), but not obligatorily so.

We show that this framework makes systematic quantitative and qualitative cross-linguistic comparison possible. Preliminary results show, for instance, that French clefts more frequently have an anaphoric clefted element than English clefts, which is reflected in cross-linguistic differences in discourse functions (Bourgoin 2017, Hasselgård 2014).

References

Avanzi, M. 2011. Note sur le prosodie des clivées du type c’est X qu- V en français parlé. In G. Corminboeuf & M.-J. Béguelin (eds). *Du système linguistique aux actions langagières. Hommages à Alain Berrendonner*. Bruxelles: De Boeck/Duculot, 113–124.

Bourgoin, C. 2017. **The Role of the English It-Cleft and the French C'est-Cleft in Research Discourse** *Discours : Revue de linguistique, psycholinguistique et informatique* 21.

Collins, P. 2006. It-clefts and wh-clefts: Prosody and pragmatics. *Journal of Pragmatics* 38(10): 1706–1720.

Delin, J. 1990. A multi-level account of cleft constructions in discourse. *Proceedings of the 13th Conference on Computational Linguistics.* 2: 83–88.

Doetjes, J., Rebuschi, G., Rialland, A. 2004. Cleft Sentences. In F. Corblin & H. De Swart (Eds.), *Handbook of French semantics*. Stanford: CSLI Publications. 529-552.

Filppula, M. 2009. The rise of it -clefting in English: areal-typological and contact-linguistic considerations. *English Language and Linguistics* 13(2): 267–293.

Halliday, M.A.K., & Greaves, W. 2008. *Intonation in the grammar of English*. London: Equinox Pub.

Hasselgård, H. 2014. It-clefts in English L1 and L2 academic writing: the case of Norwegian Learners. In Davidse, K., Gentens, C., Ghesquière, L., Vandelanotte, L. (Eds.) *Corpus interrogation and grammatical patterns*. Amsterdam: John Benjamins. 295–319.

Kaltenböck, G. 2005. It-extraposition in English: a functional view. Int. J. Corpus Linguist. 10 (2), 119–159.

Mertens, P. 2012, La prosodie des clivées. In S. Caddéo et al. (Eds.), *Penser les langues avec Claire Blanche-Benveniste*. Aix-en-Provence: Presses Universitaires de Provence. 127-139

Prince, E. 1978. A comparison *of wh*-clefts and *it*-clefts in discourse. *Language* 54, 883–906.

Čermáková, Anna and Lenka Fárová

Reporting verbs in English, Czech and Finnish: *said* and beyond

English reporting verbs and their translation into different languages have been recently explored in several studies (e.g. Corness 2010, Fárová 2016, Čermáková & Mahlberg 2018, Nádvorníková 2020). These studies focus on *said,* the dominant reporting verb in English and explore the degree of variation when translated into other languages. The studies have shown that translations of the English verb *said* do not rely solely on the core translation equivalent in the respective languages but translations show a tendency for a higher degree of variation. Other studies on reporting verbs have also stressed the potential of reporting verbs generally as a characterising device (Ruano San Segundo 2017, Mastropierro 2020). However, all these studies treat the verb *said* as a neutral reporting verb. In this study, we wish to argue that *said* in English is more frequently than not further modified or accompanied by further descriptions, e.g. of the manner of speaking, and the verb thus is no longer neutral, as in

*"Good," she **said firmly.***

We examine these occurrences and their corresponding translations in Czech and Finnish. We use a small parallel corpus of three novels: *The Silkworm* by R. Galbraith (2014), *An Artist of the Floating Day* by K. Ishiguro (1986) and *On Beauty* by Zadie Smith (2005) and their translations into Czech and Finnish. The corpus is available online from www.korpus.cz.

The preliminary results show that with all three authors, the verb *said* occurs substantially more frequently (1.3 times more frequently for Galbraith, 1.8 times for Ishiguro and 3.1 times for Smith) accompanied by a further specification rather than as a plain, neutral, reporting verb. We have identified several recurrent types of specifications, which we have further divided into several groups based on their form. The most frequent are modifications by an adverb (usually of manner or time) as in the example above, *-ed* or *-ing* participles and prepositional phrases, as in

*"All right," he **said, defeated**.*
*"Now, you shut up," **said** Zora **beaming**.*
*"Very well," I **said, with a laugh**.*

We have examined these occurrences and their translations further with the following research questions in mind:

1. What lexico-grammatical patterns accompany reporting verb *said*?
2. How are these patterns manifested in translation?
3. What do the patterns say about the individual style of authors and/or translators?

We have identified eight major recurring lexico-grammatical patterns complementing the verb *said*. We have analysed the translations in terms of their congruence, the reporting verb that is used and possible shifts, and corresponding lexico-grammatical structures in the target languages. Our results clearly point to author and translator preferences, for example, while the prepositional phrase complementation relying on the preposition *with* (see example above) is similarly frequent with Galbraith and Ishiguro, it is less preferred by Smith. In *with* prepositional clauses, Galbraith indicates both the manner of speaking and body language accompanying the speech, Ishiguro shows preference for body language descriptions and Smith for manner. In translation of these, Finnish tends to use the main translation equivalent *sanoi* while Czech uses a greater variety of verbs. In the translations of the modifying elements, translator preferences seem to be manifested.

References:

Corness, P. 2010. Shifts in Czech Translations of the Reporting Verb *Said* in English Fiction. In F. Čermák, P. Corness & A. Klégr (eds), *InterCorp: Exploring a Multilingual Corpus*, pp. 159-177. Praha: Nakladatelství Lidové noviny.

Čermáková, A. & Mahlberg, M. 2018. Translating fictional characters – Alice and the Queen from the Wonderland in English and Czech. In A. Čermáková & M. Mahlberg (eds), *The Corpus Linguistics Discourse: In Honour of Wolfgang Teubert*, pp. 223-253. John Benjamins.

Fárová, L. 2016. Uvozovací slovesa v překladech třech různých jazyků. In A. Čermáková, L. Chlumská & M. Malá (eds), *Jazykové paralely*, pp. 145-161. Praha: Nakladatelství Lidové noviny.

Mastropierro, L. 2020. The translation of reporting verbs in Italian. The case of the *Harry Potter* series. *International Journal of Corpus Linguistics* 25 (3), 241-269.

Nádvorníková, O. 2020. Differences in the lexical variation of reporting verbs in French, English and Czech fiction and their impact on translation. *Languages in Contrast* 20 (2), 209-234.

Ruano San Segundo, P. 2017. Reporting verbs as a stylistic device in the creation of fictional personalities in literary texts. *Journal of the Spanish Association of Anglo-American Studies*, 39 (2), 105-124.

Egan, Thomas (Inland Norway University of Applied Sciences)

The dative alternation in English and Norwegian: verbs of SENDING, BRINGING, LENDING and SELLING

This paper presents the results of a study of double object constructions containing the cognate verbs English *send, bring, lend* and *sell* and Norwegian *sende*, *bringe*, *låne* and *selge,* using data from the English–Norwegian Parallel Corpus (ENPC: see Johansson 2007: 10). It is the third and final paper on cognate verbs in the two languages that partake of the dative alternation. The first two papers dealt with the two most common English ditransitive verbs, *give* and *tell*, labelled 'typical ditransitive verbs' by Mukherjee (2005), and their Norwegian cognates. Egan (forthcoming^1) presents an analysis of English *give* and Norwegian *gi* constructions and shows that these are remarkably similar, both in their semantics and their distribution. Egan (forthcoming^2) contains an analysis of English *tell* and Norwegian *fortelle* constructions and shows that these are very dissimilar indeed.

The reason for selecting cognate verbs for the study is grounded in the assumption that translators, in addition to attempting to render the semantic and pragmatic import of their source texts, will tend to employ congruent syntactic constructions where these are available in the target language (see Ebeling 1998: 169). We may expect this point to apply *a fortiori* to constructions that are not only syntactically, but also lexically congruent in the sense that they contain a cognate verb. Given this premise, an analysis of the translation correspondences of the four pairs of constructions may be expected to throw further light on the similarities and differences between their distributions in the two languages.

In one of the four pairs of verbs in the present study, that of the SEND verbs, the English member, *send,* is labelled a 'habitual ditransitive verb' by Mukherjee (2005), signifying that it is regularly employed with two objects, but not nearly as often as *give* and *tell*. The other three English verbs are classed as 'peripheral ditransitive verbs', since they only occur occasionally in double object constructions. A first look at the occurrences of all eight verbs in the ENPC suggests that this distinction may also be applied appropriately to the four Norwegian verbs.

Three research questions are posed for each of the four pairs of verbs in the study

1. How similar to/different from one another are the distributions of the ditransitive and prepositional constructions containing the English and Norwegian verbs in the original texts in the two languages?
2. Are there some kinds of tokens that are usually, or seldom, translated by congruent constructions? What characterises these?
3. What characterises translations that are divergent in form?

Although none of the four pairs of verbs in the study require a concrete THEME as direct object (one can send someone a thought or lend them an ear, for example, or sell them an intellectual property), the majority of examples of all four do encode the physical transfer of a concrete object. One might therefore hypothesise that they will bear a close resemblance in their distribution to GIVE. It turns out that this is the case for some, but not all, of the verbs. An explanation will be proposed for the difference(s) between them.

References

Egan, Thomas. forthcoming^1. Giving in English and Norwegian: a contrastive perspective. In Melanie Röthlisberger, Eva Zehentner and Timothy Colleman (eds), *Ditransitive Constructions in Germanic Languages*. Amsterdam: John Benjamins.

Egan, Thomas. forthcoming^2. Telling in English, Norwegian and French: a three-way contrast. In Anna Cermakova, Signe Oksefjell Ebeling, Magnus Levin and Jenny Ström Herold (eds), *Crossing the Borders, Complex Contrastive Data and the Next Generation,* a special edition of *Bergen Language and Linguistic Studies.*

Ebeling, Jarle. 1998. Using translations to explore construction meaning in English and Norwegian. In Stig Johansson and Signe Oksefjell (eds), *Corpora and Cross-linguistic research: Theory, method and Case Studies.* Amsterdam: Rodopi, 169–195.

Johansson, Stig. 2007. *Seeing through Multilingual Corpora: On the use of corpora in contrastive studies*. Amsterdam: John Benjamins.

Mukherjee, Joybrato. 2005. *English Ditransitive Verbs: Aspects of Theory, Description and a usage-based Model.* Amsterdam: Rodopi.

Hasselgård, Hilde (University of Oslo)

Periphrastic genitive constructions in English and Norwegian

Both English and Norwegian can alternate between an *s*-genitive and a periphrastic genitive in the form of a postmodifying prepositional phrase as shown in (1) and (2). Where English uses *of* to express a possessive relationship, Norwegian uses the preposition *til* ('to'). The s-genitive differs only in that the Norwegian suffix -*s* is (normally) not accompanied by an apostrophe.

(1) Broren *til Tora* sykler med blomsterpakker… (BV2)
Lit: "The brother *of Tora* cycles with flower-parcels"
Tora's brother delivers flowers by bicycle. (BV2T)

(2) … he was asked what *the principal crop of Thailand* was… (JB1)
…han ble spurt hva som var *Thailands viktigste jordbruksprodukt*… (JB1T)
Lit: "he was asked what which was Thailand's principal crop"

The two languages thus possess near identical morphosyntactic resources for genitive constructions. However, "formal similarity is no guarantee that there is identity of use" (Johansson 2012, 47). Indeed, a previous study indicates that the conventions governing the alternation are not the same in the two languages (Hasselgård 2021).

Using the English-Norwegian Parallel Corpus (ENPC), the present study takes periphrastic genitives as a starting point, delimiting the data to those that express possession (Keizer 2007, 63, Mac Donald 1985, 5). The genitive alternation is expected to show through the translation correspondences, which can highlight similarities and differences between the languages. The study includes a comparison of fiction and non-fiction, as the genitive alternation is described as sensitive to register and formality (Biber et al. 1999, 302; Holmes and Enger 2018, 49). Heller et al. (2017) list possessor animacy, constituent length, and final sibilancy of the possessor as important factors in the alternation in English. The same factors may be relevant in Norwegian, though in different ways (Holmes and Enger 2018). For instance, the semantics of the possessor, especially +/- human, seem more important in English than in Norwegian (Johannessen et al. 2014).

A small pilot study of 100 translation pairs from the fiction part of the ENPC included examples where the source and translation both used the periphrastic genitive, as in (3).

(3) …the voices *of* my spirit companions… (BO1)
…stemmene *til* mine følgesvenner i åndeverdenen… (BO1T)
Lit: "…the voices to my companions in the spirit-world"

However, most translations of possessive PPs were non-congruent, often taking the form of s-genitives (in both directions of translation). This is especially apparent in the case of family relations (example (1)) where Norwegian seems to use the periphrastic genitive more than English (Johannessen et al. 2014). Other types of non-congruent translations include premodifiers, as in *the music of gods – guddommelig musikk ('god-like music'),* non-PP postmodifiers, as in *grunnen til Johns familie* ('the ground of John's family') – *land belonging to John's family* (TB1), and PPs without possessive meaning, as in *people of the city – menneskene i byen* ('the people in the city'). There seem to be individual differences between corpus texts in the use of genitive constructions, which may be due to topic and/or author/translator preferences. This will be investigated, along with the possibility of some constructions being used recurrently as phraseological chunks.

In short, the present study crosses boundaries between languages, registers, originals vs. translations, writer preferences, and morphosyntactic expressions of the 'same' relation within languages. In doing so, it takes account of the semantics and phraseology of the genitive construction in both languages.

References

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Hasselgård, Hilde. Forthcoming (2021). Lexicogrammar through colligation: Noun + Preposition in English and Norwegian. To appear in *Crossing the Borders: Complex Contrastive Data and the Next Generation*, ed. by Anna Čermáková, Signe Oksefjell Ebeling, Magnus Levin, Jenny Ström Herold. *Bergen Language and Linguistics Studies (BeLLS)* https://bells.uib.no/

Heller, Benedikt, Benedikt Szmrecsanyi, and Jason Grafmiller. 2017. Stability and Fluidity in Syntactic Variation World-wide: The Genitive Alternation across Varieties of English. *Journal of English Linguistics* 45, no. 1: 3–27.

Holmes, Philip and Hans-Olav Enger. 2018. *Norwegian. A Comprehensive Grammar*. London / New York: Routledge.

Johannessen, Janne Bondi, Marit Julien and Helge Lødrup. 2014. Preposisjoner og eierskapsrelasjoner: et menneskesentrert hierarki. In *Språk i Norge og nabolanda,* ed. by Kristin Hagen and Janne Bondi Johannessen, 65-97. Oslo: Novus forlag.

Johansson, Stig. 2012. Cross-linguistic perspectives. In Merja Kytö (ed.) *English Corpus Linguistics: Crossing Paths*, 45-68. Amsterdam: Rodopi.

Keizer, Evelien. 2007. *The English Noun Phrase. The Nature of Linguistic Categorization.* Cambridge: Cambridge University Press.

Mac Donald, Kirsti. 1985. Verden ligger for føttene til Liv Ullmann. Om preposisjoner som alternativ til genitiv. *NOA – Norsk som andrespråk*, Vol 1, 1-15.

Martinková, Michaela (michaela.martinkova@upol.cz), Markéta Janebová (marketa.janebova@upol.cz), Palacký University, Olomouc

Coronavirus pandemic across languages and waves: Evidence from newspaper-based corpora of English and Czech (WiP)

Framing discourse about diseases and/or emotional implications thereof have received considerable attention (e.g. [1], [2]), and so has discourse about epidemics ([3], [4], [5]). This presentation targets the current coronavirus pandemic and the figurative language used for its description in English and Czech newspaper texts. Since comparability is hard to achieve in this type of genre [6], only aspects relevant for the analysis were considered in the process of corpus creation: newspaper section (Top Stories, Coronavirus, National, World and Financial sections of *The Guardian* [TGC] and respective sections of *Lidové noviny* [LNC]) and time frame (09/2020-02/2021, i.e., the second wave of the pandemic). The corpora were created in Sketch Engine (SkE) and explored with SkE tools.

The Keyword tool ([7]) was used to generate lists of expressions over-represented in the two corpora. Candidates for metaphorical expressions were identified and compared cross-linguistically, along with their collocation candidates. Metaphorical collocation candidates were also identified for the expressions *Covid(-19)*, *(corona)virus*, *pandemic,* and *epidemic* and their Czech counterparts and their over-representation was tested with SIGIL: Corpus Frequency Test Wizard.

Preliminary results reveal more similarities than differences. Both languages make use of CONTAINER schemas for the beginning and spreading of the pandemic (*outbreak*/*vypuknutí*) and for the governmental responses: *closure*/*uzávěra* are among keywords, *lockdown* even in both languages (in LNC first attested on October 10). Orientational metaphors are used for keeping track of coronavirus cases (*rise, nárůst* [growth]) as well as for the (economic) impact (*fallout*, *propad* [drop]). Structural metaphors include the conceptualizations of the (corona)virus, Covid-19 and the epidemic/pandemic (often metonymically used) as natural disasters (*worst/hard-hit*, *zasažený* [hit]), most typically as FIRE (*firebreak, ohnisko* [centre of fire]) or WATER (*wave/vlna*). In LNC, also the verb z*krotit* [to tame] features among keywords, rendering Covid as a BEAST, as well as verbs with negative semantic prosody (*polevovat* [abate], *bujet* [grow rampant]). As to WAR metaphors, though only *frontline*, *fáze boje* [phase of fight] and *nálož* [load] are among SkE keywords, words from the WAR domain collocate with words for the virus, disease or pandemic, and the Czech noun *boj* [fight] is over-represented in the LNC with respect to Czech_Web_2017_Sample at p<0.001. Differences include reference to deaths: in Czech those who die of coronavirus are its "victims" (*oběti*) where English has *death toll.* Unlike English, Czech lexicalizes the process leading to the attainment of *herd immunity*, *promořování*. Though metaphorically related to the process of "staining wood", the use of the word by authorities was not well accepted for its homonymy with a panslavic word cognate to the Czech noun meaning "plague".

In the next stage, a corpus of Czech texts covering the first wave of the epidemic will be created and the use of military metaphors compared. Presumably, the change in the public attitude towards the restraining measures observed between the first and second wave in the Czech Republic will allow us to test the hypothesis that WAR metaphors are "ill-equipped to make people abstain from their usual behaviours" (Semino quoted in [5], see also [8]).

References

[1] Hendricks, Rose K., Zsófia Demjén, Elena Semino, and Lera Boroditsky. 2018. Emotional implications of metaphor: Consequences of metaphor framing for mindset about cancer. *Metaphor and Symbol* 33, no. 4: 267-279.

[2] Semino, Elena, Zsófia Demjén, Andrew Hardie, Sheila Payne, and Paul Rayson. 2018. *Metaphor, Cancer and the End of Life: A Corpus-Based Study.* London, UK: Routledge.

[3] Wallis, Patrick, and Brigitte Nerlich. 2005. Disease metaphors in new epidemics: the UK media framing of the 2003 SARS epidemic. *Social Science & Medicine* 60.

[4] Wicke, Philipp, and Marianna M. Bolognesi. 2020. Framing COVID-19: How we conceptualize and discuss the pandemic on Twitter. Available at
https://doi.org/10.1371/journal.pone.0240010.

[5] Beyond the battle, far from the frontline: a call for alternative ways of talking about Covid-19. Available at https://www.lancaster.ac.uk/linguistics/news/beyond-the-battle-far-from-the-frontline-a-call-for-alternative-ways-of-talking-about-covid-19.

[6] Kennin, Marie-Madeleine. 2010. What are parallel and comparable corpora and how can we use them? In A. O'Keeffe and M. McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics*. London: Routledge

[7] Kilgarriff, Adam. 2009. Simple maths for keywords. In *Proceedings of Corpus Linguistics Conference CL2009*, Mahlberg, M., González-Díaz, V. & Smith, C. (eds.). University of Liverpool, UK.

[8] Hauser, David and Norbert Schwartz. 2015. The War on Prevention. *Personality and Social Psychology Bulletin* 41:66-77. Available at
https://www.researchgate.net/publication/267742043_The_War_on_Prevention

Oksefjell Ebeling, Signe (University of Oslo)

Seeing through languages and registers: A closer look at the cognates *see* and *se*

Inspired by previous research on the English-Norwegian verb pair *see* and *se* (Øhman 2006; Ebeling & Ebeling 2020), this paper aims to shed further light on these verbs in a cross-linguistic comparison across registers. Despite being cognates, referring to the situation of perceiving with one's eyes, they have developed divergent polysemies (Aijmer 2004), i.e. meanings that do not (fully) overlap. Indeed, Øhman (2006) suggests that English *see* is more commonly used in Material processes than Norwegian *se*, as in example (1) where the Material process of meeting someone is translated into *treffe* 'meet'. Using *se* in this context would not be considered idiomatic Norwegian, as *se* rather carries the meaning of Mental perception.

(1) But then, that was the only man he had allowed her to *see*, ... [ENPC/GN1]
Men han var jo den eneste mannen han hadde gitt henne lov til å *treffe*, …

In their cross-linguistic comparison of dialogue vs. narrative in the English-Norwegian Parallel Corpus (ENPC), Ebeling & Ebeling (2020) uncovered differences in the use of *see* between the two fictional sub-registers. While *see* was most commonly used in the prototypical Mental perception sense in narrative passages (2), there was a bias towards the Mental cognition sense in dialogue (3).

(2) I could hardly *see* them in the tree. [ENPC/RDO1]
(3) "I *see* your point," .... [ENPC/DL2]

Against this background of differences between the cognates both across the two languages and the two sub-registers, the current study expands the object of study to include non-translated Norwegian dialogue and narrative from the ENPC as well as another register, namely football match reports from the English-Norwegian Match Report Corpus (ENMaRC). Preliminary scrutiny of the ENPC and ENMaRC material substantiates previous findings regarding the cognates' overlapping uses and divergent polysemies. Moreover, certain uses seem to be unique to one language and/or register, e.g. the Relational use of *see* in the sense of 'have' in the English match reports.

(4) City continued to *see* their fair share of the ball … [ENMaRC/NC]

Aiming to pin down language-specific and register-specific uses of the verbs, the study starts with an overview of the distribution of the lemmas in comparable original texts in the two languages and three registers, before analysing the verb forms *see* and *se* to answer the following RQs:

- How different/similar are the cognates' lexico-grammatical behaviour?
- Is language or register more decisive for the cognates' lexico-grammatical behaviour?

References

Aijmer, K. 2004. The Interface between Perception, Evidentiality and Discourse Particle Use – Using a Translation Corpus to Study the Polysemy of SEE. *TradTerm* 10, 246–277.

Ebeling, S.O. & J. Ebeling. 2020. Dialogue vs. narrative in fiction: A cross-linguistic comparison. *Languages in Contrast* 20:2, 289–314.

Øhman, B.I. 2006. An SFG Perspective on the Polysemy of *See*: A Corpus-based Contrastive Study. Unpublished MA thesis, University of Oslo.

Rabadán, Rosa rosa.rabadan@unileon.es, Camino Gutiérrez-Lanza camino.gutierrez.lanza@unileon.es (University of León, Spain)

Corpus-based contrast in audiovisual customization (w-i-p)

The audiovisual industry needs technical, cultural and linguistic expertise to glocalize their products, drawing on many disciplines, including contrastive linguistics. In dubbing, cross-linguistic contrast is traditionally identified with visual phonetics (Fodor 1976) and lip-syncing. However, the creation of fake spontaneous conversation among characters (*pre-fabricated orality*) presents additional language-related difficulties such as the transfer of interjections and discourse markers (Baños Piñero and Chaume 2009; Baños Piñero 2013), which are often formally dissimilar across languages. Thus, linguistic customization is far from obvious. Word-for-word translation tends to be overused to favour isochrony (Whitman-Linsen 1992: 22), but it is detrimental to the recreation of pre-fabricated orality. This strategy affects acceptability, tenor, and audience engagement (Rabadán and Gutiérrez Lanza 2020) and often results in *dubbese* (Chaume 2007 and 2020; Romero Fresco 2006 and 2009) creating ineffective communication patterns (Rabadán 2008 and 2010). These practices are also typical of cartoon dubbing, and Spanish audiences have developed a high tolerance from childhood. As a result, 'third code' dubbese (Corrius 2005) has moved into children's speech and other non-audiovisual fiction genres, becoming a vacuous Spanish dialect.

This presentation reports on work-in-progress on one of the main problem-triggers in the recreation of this pre-fabricated orality: first-person uses of English modals *can* and *could*, which in the translations are devoid of the meaning functions they had in the original, becoming useless -and noisy-function words (*puedo*, *podemos*, etc.). We aim to put forward alternative lines of action, consistent with oral features and compatible with the need for isochrony and lip-syncing. To this end, we have started to compile a translated film script subcorpus as part of CETRI (Corpus del Español TRaducido del Inglés – Corpus of Spanish Translated from English). CETRI contains materials from 2010 onwards distributed into two major subcorpora: fiction, roughly 19 million words, and non-fiction, with over 9 million words (as of December 2020). Texts have been PoS-tagged, and other annotation layers are being added gradually to allow for more refined searches. Translated scripts constitute a tiny part of the fiction subcorpus, roughly 50,000 words.

For this pilot study, we have focused on the film *Jack Reacher* (McQuarrie 2012). The customized TT2 (14,824 words) has been aligned with the intermediate, uncustomized translation, TT1 (12,799 words), and with the ST script (35,066 words), using TAligner 3.0 (Gutiérrez Lanza and Alonso 2011, MINECO 2019). Data yield essential information about the origin(s) of the Spanish *poder* renderings and their meanings in the English text's pre-fabricated orality. Additionally, CETRI TT materials have been measured against original Spanish data from the equivalent script subcorpus (*guiones*) from the Real Academia contemporary reference corpus CORPES XXI (Corpus del Español del Siglo XXI – Corpus of 21st Century Spanish). Results suggest that, by applying corpus-based findings, mistransferred uses can be changed into realistic pre-fabricated orality in Spanish, greatly improving the audiovisual text's understanding. They also imply that artificial modes of orality can be curbed if better options are available.

References

ACTRES and TRACE. 2019. Corpus del Español TRaducido del Inglés (CETRI). https://actres.unileon.es/internal/general_login/?url=/herramientas/cetri Accessed 28 December 2020.

Baños Piñero, R. and Chaume, F. 2009. *Prefabricated orality. a challenge in audiovisual translation*. InTRAlinea. Special issue: *The Translation of Dialects in Multimedia*. http://www.intralinea.org/specials/article/1714

Baños Piñero, R. 2013. La compensación como estrategia para recrear la conversación espontánea en el doblaje de comedias de situación. *TRANS* 17 (2013), 71-84**.** DOI: https://doi.org/10.24310/TRANS.2013.v0i17.3228

Chaume F. 2020. Dubbing. In Bogucki Ł. and Deckert M. (eds.). *The Palgrave Handbook of Audiovisual Translation and Media Accessibility*. Palgrave Studies in Translating and Interpreting. Palgrave Macmillan, Cham. DOI: https://doi.org/10.1007/978-3-030-42105-2_6

Chaume, F. 2007. Quality Standards in Dubbing: a Proposal. *TradTerm* 13. 71-89.

Corrius, M. 2005. The Third Language: A Recurrent Textual Restriction that Translators Come across in Audiovisual Translation. *Cadernos de Traduçao, 16*, 147–160.

Fodor, I. 1976. *Film Dubbing*. Hamburg: Buske.

Gutiérrez Lanza, C. & Alonso, J. 2011. The TRACE Corpus Aligner: Developing a new electronic tool for language researchers. *III Congreso Internacional de Lingüística de Corpus. CILC 2011. Las tecnologías de la información y las comunicaciones: presente y futuro en el análisis de córpora*. Universitat Politècnica de València. 7-9 April.

McQuarrie, C. 2012. *Jack Reacher*. https://www.imdb.com/title/tt0790724/?ref_=nv_sr_srsg_0

Rabadán, R. and Gutiérrez Lanza, C. 2020. Developing Awareness of Interference Errors in Translation: An English-Spanish pilot study in popular science and audiovisual transcripts. *Lingue e Linguaggi* 40. 137-168.

Rabadán, R. 2008. Refining the Idea of 'Applied Extension.' In Pym, A.; Shlesinger, M. and Simeoni, D. (eds.). *Beyond Descriptive Translation Studies*. Amsterdam: John Benjamins. 103–117.

Rabadán, R. 2010. Applied Translation Studies. In Gambier, Y. and van Doorslaer, L. (eds.). *Handbook of Translation Studies*. Volume 1. Amsterdam: John Benjamins. 7–11.

Real Academia de la Lengua Española. 2019. Corpus del español del siglo XXI (CORPES XXI). https://www.rae.es/recursos/banco-de-datos/corpes-xxi Accessed 28 December 2020.

Romero Fresco, P. 2006. The Spanish Dubbese: A case of (un)idiomatic Friends. *The Journal of Specialised Translation* 6. 134-151.

Romero Fresco, P. 2009. Naturalness in the Spanish Dubbing Language: A case of not-so-close Friends. *Meta*, 54/1. 49-72.

MINECO. 2019. CorpusNet. TAligner: http://corpusnet.unileon.es/herramientas-tecnicas Accessed 28 December 2020.

Whitman-Linsen, C. 1992. *Through the Dubbing Glass: The synchronization of American motion pictures into German, French, and Spanish*, Peter Lang, Frankfurt am Main and New York.

Ström Herold, Jenny and Magnus Levin (Linnaeus University)

From dashes to dashes? – a contrastive corpus study of dashes in English, German and Swedish

Our paper investigates dashes from an English-German-Swedish perspective, focusing on their frequencies and distributions in originals and translations. In translation studies, punctuation has been largely overlooked, which is surprising as the appropriate use of punctuation marks is no trivial matter for translators (Ingo 2007: 67; Shiyab 2017: 93−101). The present study continues our series of contributions on punctuation in contrast, which so far have addressed colons (Ström Herold & Levin forthcoming) and brackets (Levin & Ström Herold forthcoming).

Our material comprising 13,000 dashes and 6,000 "non-dash" correspondences stems from the Linnaeus University English-German-Swedish corpus (LEGS) (Ström Herold & Levin forthcoming), which contains non-fiction texts from the 2000s. A pilot study reveals that dashes partly serve similar functions to those for brackets in Levin & Ström Herold (forthcoming).

(1) Folklore is full of bloodsucking, flesh-eating monsters – *vampires, ghouls, ogres and werewolves* – which feed on humans as they sleep.
(2) […] attitudes to affairs – *at least amongst the upper classes* – were generally tolerant, […]
(3) To start with we used a GPS to map each bee's position – *that was until I accidentally left the GPS on the roof of the car and drove off*.

The functions range from marking off short exemplifying elaborations (as illustrated in (1)), to hedges (in (2)) and longer, clausal specifications (as in (3)) where the dash symbolizes a break, preparing readers for something unexpected (cf. Duden RgD 1997: 292). Similar to brackets, dashes thus seem to serve either content-elaborating or more reader-oriented functions.

The findings in original texts show that German writers use considerably more dashes than English and Swedish ones. German and Swedish prefer the sentence-final position (as illustrated in (4) below), while English uses medial (as in (1) and (2)) and final position equally frequently. In translations, two major trends emerge: i) English translators "add" the most brackets, and Swedish ones the least, and ii), there are striking differences in the proportions of the brackets retained by translators. The frequencies of additions can be explained by English being the most "powerful" language and Swedish the least (Levin & Ström Herold forthcoming). The proportions of retained brackets do not adhere to such neat patterns, ranging from 85% retained to less than half. As translators typically retain large majorities of all punctuation (Frankenberg-Garcia 2019), our paper will explore these differences closely.

Similar to our brackets study (Levin & Ström Herold forthcoming), the most commonly used "non-dash" correspondents are commas, irrespective of translation direction, as in (4):

(4) Vill du vara säker – *använd stektermometer*. (Swedish original)
If you want to be sure, *use a thermometer*.

Interestingly, it has been suggested that dashes – compared to both commas and brackets – are much less formal and have a "dramatic flair" (Quirk et al. 1985: 1629; Crystal 2015: 158), indicating certain stylistic differences. Thus, with this trilingual data set-up, we will address forms, functions and stylistic concerns while disentangling language-specific preferences and translation-induced changes.

References

Crystal, David. 2015. *Making a point: the pernickety story of English punctuation*. London: Profile Books.

Duden. Band 9. *Richtiges und gutes Deutsch (RgD). Wörterbuch der sprachlichen Zweifelsfälle*. 1997. Berlin: Dudenverlag.

Frankenberg-Garcia, Ana. 2019. A *corpus study* of splitting and joining sentences in translation. *Corpora* 14(1), 1–30.

Ingo, Rune. 2007. *Konsten att översätta. Översättandets praktik och didaktik*. Lund: Studentlitteratur.

Levin, Magnus & Jenny Ström Herold. Forthcoming. On brackets in translation (or how to elaborate in brackets). To appear in *Bergen Language and Linguistics Studies* (*BeLLS*).

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, & Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London: Longman.

Shiyab, Said M. 2017. *Translation: Concepts and critical issues*. Antwerp: Garant Publishers.

Ström Herold, Jenny & Magnus Levin. Forthcoming. The colon in English, German and Swedish: a contrastive corpus-based study. To appear in *Comparative Punctuation*. *Linguistik – Impulse* & Tendenzen. Berlin, New York: Walter *de Gruyter*

Workshop 3: Exploring Powerful Tools to Ensure Robust and Reproducible Results in Corpus Linguistics

Convenors: Martin Schweinberger (The University of Queensland, Australia), Joseph Flanaghan (University of Helsinki), Gerold Schneider (University of Zurich)

This workshop explores how powerful tools enable researchers to come up with efficient workflows and pipelines which allow them to stand on the shoulders of giants and at the same time produce robust and replicable results. Fully scripted workflows, such as R or Python scripts have the advantage that results can be updated, reproduced, and shared at the push of a button (Flanagan 2017). Markdown and Text can be integrated to offer a seamless transition between data and publication. Powerful user-friendly tools such as LancsBox, LightSide, AntConc or VARD can be installed by every user to apply state-of-the-art approaches to obtain reproducible results in few, well-defined steps. The advantages of these tools are that they have the potential to counteract the increasing loss of public trust in research from the Humanities and Social Science (Yong 2018). Advanced statistical approaches on the one hand offer new insights, easing the step from data to evidence (Suhr et al. 2019, Schneider et al. 2017), for example by offering higher levels of robustness, such as cross-validation, regulation, overfit warnings and built-in evaluation. But on the other hand, some may also introduce new challenges to reproducibility and robustness: Topic Modelling and sampling may depend on random seeds, seemingly similar parameters can lead to strongly different results across similar or even the same tool –how should we deal with them? As such, the workshop advances the discussion about Best Practices in (corpus) linguistics (Berez-Kroeker et al. 2018) and aims to raise awareness about existing resources and problematizes practices in (corpus) linguistics that hinder transparency, replicability, and high quality of research outputs. The workshop proposes approaches and invites contributions transparency and high quality of research in (corpus) linguistics, as well as publication practices (pre-registration, open access, and pre-prints). Specifically, the workshop addresses the following issues:

(i) How can the FAIR principles (Findable, Accessible, Interoperable, and Reusable) be observed?; (ii) How can transparency and replicability of research be enhanced by using collaborative tools(e.g. Google Docs, Git, Docker, shiny R)?;

(iii) How can Jupyter or R Notebookshelpto document analyses and making them available to the community and reviewersenable full reproducibility?;

(iv) How can researchers profit the most from integrating user-friendly out-of-the-shelf applications in their workflows?

(v) What are advantages and disadvantages of powerful blackbox methods (e.g. BERT) and or as opposed to simpleclose-to-text methods (e.g. Concordancing)?

(vi) Should we aim for robustness or reproducibility?

(vii) How can we document workflows and prevent data loss or corruption?

The workshop format will include detailed recommendations on practices and tools, presentations of submitted research contributions and a substantial open panel discussion.

Haugh, Michael (The University of Queensland) and Simon Musgrave (Monash University)

Looking for a good laugh: A combinatorial approach to identifying pragmatic phenomena in spoken corpora

Although working with data from spoken corpora holds considerable promise for pragmatics, one of the challenges facing those wishing to analyse the interactional practices by which we structure and make sense of communicative interaction is how to assemble a sufficiently large dataset of the practice in question given these practices are context-sensitive. Many pragmatic phenomena cannot be readily identified in spoken corpora because what a particular linguistic form is taken to mean is not simply a function of its composition, and the genre or activity type in which it appears, but is also determined, in part, by its position within and across a sequence of utterances (Schegloff 1993). Standard concordance-based approaches can therefore only get us so far. What is needed is the ability to search for particular *forms* in particular sequential *positions* across different situated contexts (Rühlemann and Gee 2017; Rühlemann 2018). In short, identifying pragmatic phenomena in spoken corpora requires the implementation of combinations of features of form and position, or what we here term combinatorial search (Haugh and Musgrave 2019).

Laughter is a case in point. Early work in conversation analysis demonstrated that laughter does not simply "flood out", but is produced in orderly ways (Jefferson 1979, 1985). What laughter is taken to be doing depends on characteristics of its form (e.g. soft breathy laughter versus loud distinct laugh particles) and its position (e.g. turn-initial versus turn-final). In this paper, we outline a prototype that employs a series of R-based scripts to identify different forms of laughter in different sequential positions in spoken corpora. We begin by first outlining the ontology we developed to account for different forms and positions of laughter. We then describe how this ontology was operationalised through a series of interlinked scripts in an R notebook. Our aim is to show it is possible to reliably identify pragmatic phenomena in spoken corpora (i.e. that the search procedure is robust), and that this procedure can be extended to identifying other examples of the phenomena in question across different spoken corpora (i.e. that the search procedure is reproducible). We argue that the key to ensuring robust and reproducible identification and analysis of pragmatic phenomena in spoken corpora is transparency with respect to both the ontology one employs and the search procedures themselves.

References

Jefferson, G. (1979). A technique for inviting laughter and its subsequent acceptance-declination. In G. Psathas, ed., *Everyday Language. Studies in Ethnomethodology*. New York: Irvington, pp. 79-95.

Jefferson, G. (1985). An exercise in the transcription and analysis of laughter. In T. van Dijk, ed., *Handbook of Discourse Analysis. Volume 3: Discourse and Dialogue*. London: Academic Press, pp. 25-34.

Haugh, M. and Musgrave, S. (2019). Conversational lapses and laughter: Towards a combinatorial approach to building collections in conversation analysis. *Journal of Pragmatics* 143: 279-291.

Rühlemann, C. and Gee, M. (2017). Conversation analysis and the XML method. *Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion* 18: 274-296.

Rühlemann, C. (2018). *Corpus Linguistics for Pragmatics.* London: Routledge.

Schegloff, E. (1993). Reflections on quantification in the study of conversation. *Research on Language and Social Interaction,* 26, 99-1

Reveilhac, Maud (Université de Lausanne) & Gerold Schneider (University of Zurich)

Detecting Stance from Corpora of Short Texts

We propose and evaluate a linguistic approach to sentiment analysis, building on rule-based approaches (Neviarouskaya et al. 2009) to which we add corpus linguistic stylistics and patterns that express stance, while also profiting from state-of-the-art computational linguistic studies on sentiment analysis.

Sentiment analysis commonly refers to determining one's stance towards a target, the valence or polarity of a text (Young & Soroka 2012), but also to the detection of emotions . The vast majority of valence and emotion classification approaches employ machine learning (ML) (e.g. AlDavel & Magdy 2020))that learns a model from training instances labelled with the correct sentiment to predict sentiment of previously unseen instances. Alternatively, lexicon-based approaches are also frequent. They use lists of words associated with valence and emotion categories (Pennebaker et al. 2015).

There is a clear lack of linguistically motivated approaches to stance detection. As RQ 1, we hypothesize that our proposed approach with a focus on well-studied linguistic features and patterns, such as nominalisation, passivisation, hedges, verb/noun ratio, use of rare words and pronouns can play a complementary role in ML and dictionary approaches, especially for the detection of stance in short messages, which are often written in abbreviated or expressive styles and where data sparseness is acute. Using SemEval 2016 data as a train set (Mohammad et al. 2017) and IBM ClaimStanceDataset as test set (Bar-Haim et al. 2017), we also combine conditional inference trees with the analysis of random forests to investigate the interplay between linguistic features in stance recognition with stance-valence combinations.

To detect further relevant syntactic patterns, we test each sentence for valence and emotion using existing dictionaries as seeds, and then use a syntactic dependency parser to semi-automatically detect syntactic patterns that express stance, add rules concerning modifiers and entity recognition, and estimate the overall stance of each sentence. We evaluate our classification accuracy and compare our model to ML approaches with bag-of-words models. As RQ 2, we hypothesize that our model is linguistically better interpretable, more robust and may perform better, because ML models tend to "overfit" to the training data, thus poorly predicting data from different domains or periods. We document and script each processing step for reproducibility and replicability, and discuss robustness (https://coderefinery.github.io/reproducible-research/01-motivation/).

References:

Neviarouskaya, A., Prendinger, H., & Ishizuka, M. (2009). "Compositionality principle in recognition of fine-grained emotions from text". In *Third International AAAI Conference on Weblogs and Social Media*.

Young, L., & Soroka, S. (2012). "Affective news: The automated coding of sentiment in political texts". *Political Communication*, 29(2), 205-231.

AlDayel, A., & Magdy, W. (2020). "Stance Detection on Social Media: State of the Art and Trends". arXiv preprint 2006.03644.

Pennebaker, J. W., Boyd, R., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. University of Texas at Austin. https://doi.org/10.15781/T29G6Z

Mohammad, S. M., Sobhani, P., & Kiritchenko, S. (2017). "Stance and Sentiment in Tweets". *Special Section of the ACM Transactions on Internet Technology on Argumentation in Social Media*, 17(3).

Bar-Haim, R., Edelstein, L., Jochim, L., & Slonim, N. (2017). "Improving Claim Stance Classification with Lexical Knowledge Expansion and Context Utilization". In *Proceedings of the 4th Workshop on Argument Mining*, 32-38.

Schneider, Gerold (University of Zürich)

Text Crunching Center (TCC): Data-driven methods for linguists, social science and digital humanities

This talk introduces the Text Crunching Centre (TCC) which is a Computational Linguistics and Digital Humanities service hosted at the University of Zurich, and a collaboration partner of LADAL. We present teasers from text analytics involving social, political and historical studies and from coaching sessions with R, helping users to script their studies.

Schweinberger, Martin (UQ, UiT)

The Language Technology and Data Analysis Laboratory (LADAL) - building computational humanities infrastructures: experiences, problems, and potentials

This presentation introduces the *Language Technology and Data Analysis Laboratory* (LADAL), and discusses the implications of our experiences to date in establishing it for broader efforts to develop researcher capacity in the digital humanities.

While computers are now part of every domain of life and offer a vast potential for humanities research, remarkably few resources for training and upskilling are available to researchers in the humanities, arts, and social sciences (HASS). Furthermore, the replication Crisis which is an ongoing methodological crisis in the life sciences, has raised the awareness for the necessity for transparent and reproducible research practices. The LADAL represents an effort to provide a resource that addresses the gap in systematic training in data processing, visualization, and analytics as well as data management.

As such, the LADAL represents a school-based computational humanities resource infrastructure maintained by the School of Languages and Cultures at the University of Queensland. It aims to assist staff and postgraduate students within the UQ School of Languages and Cultures to learn how to use data analytics, digital research tools, and other forms of technology to enhance their existing research programs, as well as offer pathways to new research possibilities. It complements the more generic resources and training in digital humanities methods offered by libraries (e.g., the Digital Scholars Hub at UQ) with the more specialised training/support in particular digital research methods and technologies that are required by researchers working on specific languages and cultures.

The LADAL consists of a specialist computing lab for language-based computational and experimental work (the Computational and Experimental Workshop) and an online virtual lab. With respect to web-based materials, the LADAL website (https://slcladal.github.io/index.html) offers self-guided study materials and hands-on tutorials on topics relating to digital tools, computational methods for data extraction and processing, data visualization, statistical analyses of language data, and provides links to further resources and short descriptions of digital tools relevant for digital HASS research.

In addition, the LADAL offers face-to-face consultations and specialized workshops. UQ researchers are encouraged to contact LADAL staff for advice and guidance on matters relating to digital research tools, data visualization, various statistical procedures, and text analytics. The talk will also provide information about its relations to the Australian Text Analytics Platform (ATAP) which represents an effort to promote text analytics in Australia and to make resources for using text analytics available to a wider community of researchers.

Staff feedback during face-to-face consultations and workshop attendance confirms there is substantial demand for the kind of digital humanities infrastructure offered by LADAL. It also suggests that support and training for researchers in the digital humanities should be conceptualised on a continuum from more generic through to more localised support.

Wallis, Sean

How do we understand significance?

Statistical significance is the key concept on which inferential statistics is based, but it is counterintuitive and frequently poorly understood. The most likely reason is cognitive: lived human experience is of serial single episodic events, whereas statistical inference concerns the variation in many observed events (Wallis 2021, Rafi and Greenland 2020). Statistical measurement is obtained through counting and recollection, and concepts of probability and variation are intangible to us. The well-known 'belief in small numbers' (Tversky and Kahneman 1971) is a failure to estimate the (un)certainty of observed results. With no ability for researchers to cross-check results, the output of tools and methods is taken on trust.

Common discipline-independent errors include confusing scatter intervals and confidence intervals (and replication intervals), the Normal fallacy and citing Gaussian 'standard error' for proportions (and derived functions of proportions), failure to engage continuity and other corrections, and employing 'p-values' to compare results (Nieuwenhuis et al. 2011). Solutions are documented in (Wallis 2012-, 2021).

A very common set of errors consists of employing a method specified by a mathematical model with data that does not conform to the model's assumptions. For example, logistic regression and log likelihood tests assume observed proportions are free to vary from 0 to 1, but corpus linguistics data with per million word baselines are commonly employed. Patently, no algorithmic improvement can substitute for a poor experimental design and absent baseline data. The ultimate problem is that 'statistics' has become disconnected from logical reasoning. Research papers stop abruptly following citation of results, fail to spot supportable claims or contain unsupported ones.

In this paper we will outline some practical steps for teaching and reporting statistical uncertainty that can begin to redress this balance. First, students should be taught about probability theory and the correct derivation of confidence intervals through inversion of resampling distributions. Second, students should be taught good research practice, to plot graphs with confidence intervals, provide clear reports of experimental design and employ meaningful baselines. They should verify their data, minimising false positives and negatives, and test instances for alternation. Fundamentally they should be taught to engage critically with data and method, and learn how to draw cautious results.

References

Nieuwenhuis, S., Forstmann, B.U. and Wagenmakers, E.-J. 2011. Erroneous analyses of interactions in neuroscience: a problem of significance. *Nature Neuroscience*, **14**, 1105-1107. DOI: 10.1038/nn.2886.

Rafi, Z. and Greenland, S. 2020. Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprise. *BMC Medical Research Methodology* **20,** 244. DOI: 10.1186/s12874-020-01105-9.

Tversky, A., and Kahneman, D. 1971. Belief in the law of small numbers. *Psychological Bulletin* **76**:2, 105-110. **»** ePublished

Wallis, S.A. 2012-. corp.ling.stats (blog). London: Survey of English Usage. https://corplingstats.wordpress.com

Wallis, S.A. 2021. *Statistics in Corpus Linguistics Research.* New York: Routledge

Weihs, Claus and Sarah Buschfeld

Variable importance in ensembles from undersampling: generalizability meets robustness

Studies in linguistics are often characterized by data sets with unbalanced class variables, which leads to models with bad balanced accuracies. Undersampling the larger class or both classes to different extents so that they have approximately the same size may solve this problem (e.g. Weihs & Buschfeld 2021a, 2021b). At the same time, undersampling allows for assessing the true predictive power of a model rather than model fit, which does not measure the generalization ability to a wider population.

In the present paper, we briefly introduce methods of undersampling for conditional inference trees (Hothorn et al. 2006). We then focus on how measuring variable importance can be incorporated in such an approach. To this end, we adapt measuring variable importance as utilized in random forests to ensembles of undersampled models. Since ensembles represent many different resamples of the original sample, their predictive power and generalizability appears to be more robust than, e.g., that of the best single tree in the ensemble.

We draw on a corpus of L1 child Singaporean and British English (Buschfeld 2020). The data from Singapore come from 30 children of different ethnicities, male and female, aged 2;5 (two years; five months) to 12;1. The data from England come from 13 monolingual and 8 bilingual children aged 2;1 to 10;9. All data were elicited by means of video-recorded task-directed dialogue between researcher and child, consisting of a grammar elicitation task, a story retelling task, elicited narratives, and free interaction. The recorded material was orthographically transcribed and manually coded for the realization of subject pronouns, i.e. realized vs. zero.

To predict the realization of subject pronouns, we model their dependence on the in-tralinguistic variables pronoun (PRN) and mean length of utterance (MLU) and a number of extralinguistic variables, i.e. ethnicity (ETH), age (AGE), sex (SEX), and linguistic background (LIBA). In order to assess predictor importance, we undersample both the large and the small class with different percentages: p-large and p-small. For each combination of p-large and p-small, we repeat the procedure 1001 times. We consider 16 different combinations of p-small and p-large. We measure variable importance for the ensemble of the three best trees of these 16 combinations, i.e. for an ensemble of 48 trees.

The results show that PRN is the most important predictor, followed by MLU. The four extralinguistic predictors AGE, LIBA, ETH, and SEX appear to be less important. For both the best individual tree and the best ensemble, the balanced accuracy is 70.3%. Thus, we have found models with comparatively high predictive accuracy that clearly depict how language use is first and foremost influenced by intralinguistic factors, but also by biological and sociolinguistic factors.

References

Buschfeld, S. 2020. *Children's English in Singapore: Acquisition, Properties, and Use*. London: Routledge.

Hothorn, T., Hornik, K., Zeileis, A. 2006. Unbiased recursive partitioning: a conditional inference framework. *J. Comput. Graph. Stat*. 15: 651–674.

Weihs, C., Buschfeld, S. 2021a. Combining Prediction and Interpretation in Decision Trees (PrInDT) – a Linguistic Example. arXiv: http://arxiv.org/abs/2103.02336.

Weihs, C., Buschfeld, S. 2021b. RePrInDT: Variation in undersampling and prediction, and ranking of predictors in ensembles for PrInDT. arXiv: forthcoming.

Workshop 4: Rescoping the theory and methodology of linguistic epicenters in World Englishes

Convenors : Pam Peters (Macquarie University, Sydney, Australia),Tobias Bernaisch (Justus Liebig University Gießen, Germany)

The most widely accepted notion of a linguistic epicenter refers to varieties that a) are Englishes in their own right, i.e. endonormatively stabilised in Schneider's (2007) nomenclature, and b) (potentially) serve as models for neighboring countries (cf. Hoffmann et al. 2011: 259). Although the concept of linguistic epicenters in the World Englishes paradigm has existed since the 1990s (cf. Leitner 1992), attention has concentrated on the evolution of individual varieties (Schneider 2007), while their interrelationships have remained on the margins of linguistic inquiry, with limited discussion in areal linguistic research (Hickey 2012) and in accounts of historical contact between closely related varieties (Millar 2016). Interest in supra-geographical contacts through the internet (Mair 2013, 2016), transnational attraction (Schneider 2014), and other extraterritorial influences (Buschfeld & Kautzsch 2020) have also diverted attention from the more localised interplay between varieties in the World Englishes paradigm. Despite recent empirical investigations into the epicentral role of Australian English for the Antipodes (Peters 2009; Peters et al. 2019), Indian English for South Asia (Hoffmann et al. 2011; Gries &Bernaisch 2016) and Singapore English for South-East Asia (Heller et al. 2017), there are many open questions in relation to epicenters and epicentral influence to be explored. Against this background, three key areas in epicentral research are in need of further academic attention and linguistic investigation, i.e. a) theory of linguistic epicenters, b) methods and evidence in epicentral research and c) linguistic epicenters in relation to other theories of variety contact such as areal linguistics, extraterritorial influence, transnational attraction or feature pools. The questions guiding the workshop are centred around these three key areas:

Theory of linguistic epicenters

- Can epicentral influence be exercised by a regional variety on its neighbors only when it has reached endonormative stabilization?
- Does epicentral influence depend on conscious recognition and acceptance of the norms of a neighboring variety –or can it occur under the radar?
- What aspects of language should be included in scoping epicentral influence –from phonology to pragmatics? Standard and nonstandard language norms?

Methods and evidence in epicentral research

- Is diachronic evidence essential for demonstrating epicentral influence?
- How can statistical analyses contribute to identifying epicentral influence?
- Do we need sociocultural and/or historical evidence to establish the contexts for epicentral influence?

Linguistic epicenters in relation to other theories of variety contact

- Could the influence of a non-adjacent hypervariety, i.e. American English, be seen as epicentral?
- How different is epicentral influence from areal contact among regional varieties?
- Does epicentral influence tend to result in convergence/levelling of linguistic features, or contribute to differentiation?

References

Buschfeld, Sarah & Kautzsch, Alexander. 2020. Modelling World Englishes: A Joint Approach to Postcolonial and Non-Postcolonial Englishes. Edinburgh: Edinburgh University Press.

Gries, Stefan Th. & Bernaisch, Tobias. 2016. 'Exploring epicentres empirically: focus on South Asian Englishes',English World-Wide37(1): 1–25.

Heller, Benedikt, Bernaisch, Tobias & Gries, Stefan Th. 2017a. ‘Empirical perspectives on two potential epicenters: the genitive alternation in Asian Englishes’, ICAME Journal 41: 111–144.

Hickey, Raymond (ed.). 2012. Areal Features of the Anglophone World.Berlin:De Gruyter Mouton.

Hoffmann, Sebastian, Hundt, Marianne & Mukherjee, Joybrato. 2011. ‘Indian English -an emerging epicentre? A pilot study on light verbs in web-derived corpora of South Asian Englishes’, Anglia129 (3–4): 258–280.

Leitner, Gerhard. 1992. ‘English as a pluricentric language’. In Michael Clyne (ed.), Pluricentric Languages: Differing Norms in Different Nations. Berlin: De Gruyter Mouton, pp. 179–237.

Mair, Christian. 2013. ‘The world system of Englishes: accounting for the transnational importance of mobile and mediated vernaculars’, English World-Wide 34(3): 253–78.

Mair, Christian. 2016. ‘Beyond and between the “Three circles”: World Englishes research in the age of globalisation’. In Elena Seoane & Cristina Suarez-Gomez (eds.), World Englishes: New Theoretical and Methodological Considerations. Amsterdam: John Benjamins, pp. 17–36

Millar, Robert. 2016. Contact: The Interaction of Closely Related Linguistic Varieties and the History of English. Edinburgh University Press. Peters, Pam. 2009. ‘Australian English as a regional epicentre’. In ThomasHoffman & Lucia Siebers (eds.), World Englishes—Problems, Properties and Prospects. Amsterdam: John Benjamins, pp. 107–124.

Peters, Pam, Smith, Adam & Bernaisch, Tobias. 2019. ‘Shared lexical innovations in Australian and New Zealand English’, Dictionaries: Journal of the Dictionary Society of North America 40(2): 1–30.

Schneider, Edgar W. 2007. Postcolonial English: Varieties around the World. Cambridge University Press.

Schneider, Edgar W. 2014. ‘New reflections on the evolutionary dynamics of World Englishes’, World Englishes 33(1): 9–32.

Götz-Lehmann, Sandra

Caught between epicenters? Tracing influences on English in Nepal

The present paper investigates India’s potential role as a linguistic epicenter for Nepal by conducting a short-term diachronic follow-up study of Bernaisch and Lange (2012) that is based on the *South Asian Varieties of English* corpus. Based on the recently updated version of the corpus with similar data from a decade later, a potential spread of the presentational *itself* construction with an adverbial focus, which has allegedly spread in Nepal originating in Indian English, is investigated. Also, possible instances of invariant use of *itself*, that are frequently used in Indian English, will be scrutinized. While findings indeed show an increase of the adverbial focus construction, supporting India’s role as an epicenter for the region, only one invariant use of the feature can be documented in Nepali English, which can tentatively be interpreted as a selection process that might be at play in the acceptance and spread of certain epicentrally induced features over others.

Gries, Stefan Th., Benedikt Heller & Tobias Bernaisch

Methods for detecting possible traces of epicentral influence: synchronic and diachronic perspectives

A linguistic epicentre can be defined as an endonormatively stabilised variety with "the potential to serve as a model of English for neighbouring countries, i.e. exert an influence on other speech communities in the region" (Hoffmann et al. 2011: 259). In order to trace said influence, one central – and currently still unresolved – question is whether the study of epicentral influence should be based on diachronic data (cf. Hundt 2013: 195) or whether synchronic investigations are reliable enough from an empirical perspective. What complicates the matter is that epicentral constellations are often assumed to exist in territories (e.g. South and South-East Asia) for which synchronic databases are available, but diachronic corpus data are sparse (despite laudable exceptions such as the continuously expanding Hansard databases; cf. e.g. Kruger et al. 2019). It is in this light that the present paper focuses on two potential epicentres, i.e. Indian English for South Asian Englishes and Singapore English for South-East Asian Englishes, to address the following research questions:

- How can the results of potential diachronic epicentral influence be detected in synchronic corpus data?
- What are the dangers associated with using synchronic data in epicentral research and – more generally – in research into postcolonial Englishes?
- How can diachronic data be used to trace potential epicentral influence?

Statistically, we rely on state-of-the-art multifactorial predication and deviation analyses with regressions/random forests (MuPDAR(F)s; cf. Gries & Adelman 2014; Deshors & Gries 2016) of the well-researched dative (*John gave Mary a book.* vs. *John gave a book to Mary.*) and genitive alternation (e.g. *the dog's owner* vs. *the owner of the dog*). The two alternations and the factors known to influence the choice between their respective variants in Asian Englishes (e.g. constituent length, animacy, etc.; cf. e.g. Bernaisch et al. 2014; Heller 2018) are analysed in synchronic corpus data for South and South- East Asian Englishes from the International Corpus of English and a newspaper corpus (cf. Bernaisch et al. 2011); the diachronic data stem from the *Historical Corpus of Singapore English* (Hoffmann et al. 2012; Hoffmann 2013).

Our results are both revealing and alarming. The synchronic data allow stating whether the structural norms of an assumed epicentre – here regarding the dative and the genitive alternation – best predict the structural choices in the other varieties of the region. While this holds true for Indian English in South Asia, it does not for Singapore English in South-East Asia. Consequently, synchronic analyses can be related to the regional model character of epicentres for a particular territory (Hoffmann et al. 2011: 259). Our diachronic study of genitives in Singapore English, however, reveals that trends deduced from synchronic corpus data are often incompatible with real-time diachronic developments, casting serious doubt on the reliability of synchronic analyses in World Englishes in general as well as with regard to epicentres. Against this background, we argue that empirically responsible identifications of epicentres be based on trends towards or diverging from an epicentre, which are derived from diachronic adjustments of the coefficients for predictors regarding given structural phenomena in the varieties assumed to be under epicentral influence.

References

Bernaisch, Tobias, Christopher Koch, Joybrato Mukherjee & Marco Schilk. 2011. *Manual for the* South Asian Varieties of English (SAVE) Corpus*: Compilation, Cleanup Process, and Details on the Individual Components*. Giessen: Justus Liebig University.

Bernaisch, Tobias, Stefan Th. Gries & Joybrato Mukherjee. 2014. "The dative alternation in South Asian English(es): modelling predictors and predicting prototypes", *English World-Wide* 35(1): 7– 31.

Deshors, Sandra C. & Stefan Th. Gries. 2016. "Profiling verb complementation constructions across New Englishes: a two-step random forests analysis of *ing* vs. *to* complements", *International Journal of Corpus Linguistics* 21(2): 192–218.

Gries, Stefan Th. & Allison S. Adelman. 2014. "Subject realization in Japanese conversation by native and non-native speakers: exemplifying a new paradigm for learner corpus research". *Yearbook of Corpus Linguistics and Pragmatics 2014: New empirical and theoretical paradigms*, ed. Jesús Romero-Trillo. Cham: Springer. 35–54.

Heller, Benedikt. 2018. *Stability and Fluidity in Syntactic Variation World-Wide: The Genitive Alternation across Varieties of English* (Unpublished doctoral dissertation). KU Leuven, Leuven, Belgium.

Hoffmann, Sebastian. 2013. "The Corpus of Historical Singapore English – Practical and methodological issues". UCREL Corpus Research Seminar. http://ucrel.lancs.ac.uk/crs/ attachments/UCRELCRS-2013-03-26-Hoffmann-Slides.pdf> (22 November 2019).

Hoffmann, Sebastian, Andrea Sand & Peter Tan. 2012. "The Corpus of Historical Singapore English – a first pilot study in data from the 1950s and 1960s". Paper presented at ICAME 33, KU Leuven, Belgium.

Hoffmann, Sebastian, Marianne Hundt & Joybrato Mukherjee. 2011. "Indian English – an emerging epicentre? A pilot study on light verbs in web-derived corpora of South Asian Englishes", *Anglia* 129(3–4): 258–280.

Hundt, Marianne. 2013. "The diversification of English: old, new and emerging epicentres". *English as a Contact Language*, eds. Daniel Schreier & Marianne Hundt. Cambridge: Cambridge University Press. 182–203.

Kruger, Haidee, Bertus van Rooy & Adam Smith. 2019. "Register change in the British and Australian Hansard (1901–2015)", *Journal of English Linguistics* 47(3): 183–220.

Hundt, Marianne, Dirk Pijpops and Laetitia van Driessche

Approaching epicentral influence with agent-based modelling: simulating factors in the choice between verb complementation patterns in speakers of Philippine and Indian English

The majority of previous empirical investigations of the epicentre hypothesis has been limited in that corpus data provide information on language use and therefore only allow to model structural similarity of varieties in synchrony (see e.g. Hoffmann et al. 2011 or Gries and Bernaisch 2016). Even diachronic corpus data would merely provide us with information on convergence or divergence (as well as more complex patterns of differential change, see Hundt 2009). One of the methodological challenges that purely corpus-based research is unable to address is the role that attitudes may play in the choice of variants in a pluricentric language like English (Hundt 2013).

Agent-based modelling provides a complementary methodology (Livet et al. 2014) that can be used to provide corpus linguists with an experimental approach to predicting variation in a speaker community (see Steels and Beuls 2013, Pijpops et al. 2015). Based on previous studies by Hundt (1998) and Pijpops and Van de Velde (2018), we provide a case study on the choice between bare-NP or PP complements for the verbs *protest* and *appeal*. We build an agent-based model (i.e. simulate language use) for two speaker communities where English is an institutionalised second language, one with American English, the other with British English as their matrilect. Our model allows for attitudes towards the two ENL varieties of English to affect the choice of verb complementation. While this approach does not provide novel insights into actual language use, it allows to gauge the effect that awareness of and attitudes towards regional varieties are likely to have in a pluricentric language, where one variety (i.e. AmE) has come to be the centre of gravity on a global scale. This insight can then be used to predict patterns of variation we should see in language use as, for instance, reflected in corpus data. We provide preliminary data from NOW to test whether the predictions are, in fact, borne out.

The results are expected to have wider implications also for the theoretical modelling of World Englishes, e.g. with Schneider's (2003, 2007) dynamic model.

References

Beuls, Katrien and Luc Steels. 2013. Agent-Based Models of Strategies for the Emergence and Evolution of Grammatical Agreement. *PLoS ONE* 8(3). e58960.

Gries, Stefan Thomas and Tobias Bernaisch. 2016. Exploring epicentres empirically: Focus on South-east Asia. *English World-Wide* 37(19): 1-25.

Hoffmann, Sebastian, Marianne Hundt and Joybrato Mukherjee. 2011. Indian English – an emerging epicentre? A pilot study on light verbs in web-derived corpora of South Asian Englishes. *Anglia* 129(3-4): 258-280.

Hundt, Marianne. 1998. *New Zealand English Grammar – Fact or Fiction? A Corpus-Based Study in Morphosyntactic Variation*. Amsterdam and Philadelphia: Benjamins.

Hundt, Marianne. 2009. Colonial lag, colonial innovation, or simply language change? In Günter Rohdenburg and Julia Schlüter, eds. *One Language, Two Grammars: Morphosyntactic Differences between British and American English*. Cambridge: University Press, 13-37.

Hundt, Marianne. 2013. The diversification of English: Old, new and emerging epicentres. In Daniel Schreier and Marianne Hundt, eds. *English as a Contact Language*. Cambridge: University Press, 182-203.

Livet, Pierre, Denis Phan and Lena Sanders. 2014. Diversity and complementarity of agent-based models in the social sciences. *Revue francaise de sociologie* 55(4): 463-500.

Pijpops, Dirk, Katrien Beuls and Freek Van de Velde. 2015. The rise of the verbal weak inflection in Germanic. An agent-based model. *Computational Linguistics in the Netherlands Journal* 5. 81–102.

Pijpops, Dirk and Freek Van de Velde. 2018. Lectal contamination. How language-external variation becomes language-internal through language contact. *Variationist Linguistics meets Contact Linguistics*. 21 May, Ascona.

Schneider, Edgar W. 2003. The dynamics of New Englishes: From identity construction to dialect birth. Language 79(2): 233-281.

Schneider, Edgar W. 2007. *Postcolonial Englishes: Varieties Around the World*. Cambridge: Cambridge University Press.

Korhonen, Minna and Adam Smith

Parliamentary Hansard as a focus for regional variance and epicentral influence in Australia, New Zealand and Papua New Guinea.

The official record of parliamentary proceedings known as Hansard is a particularly useful linguistic resource for tracking variation over time and space, as well as providing a yardstick for the degree of influence exerted by particular standards on parliamentary language in the English-speaking world. Countries throughout the Commonwealth have adopted the conventions of Hansard, first developed in the UK, and adapted them to create their own local norms. Previous research (Kruger et al., 2019) has shown how the language of the Australian House of Representatives converges with and diverges from the British House of Commons, both in terms of the editorial standards applied to the presentation of parliamentary speech (e.g. varying acceptance of contractions/split infinitives), and in the apparent choices of individual speakers (e.g. the use of *that*-complement clauses/contrasting means of expressing modality). This paper will investigate divergence between language features in Australian, British and New Zealand Hansards, using sampled diachronic corpora from the three regions covering the period 1901-2015. The BrE component has been used as a reference corpus to create a set of keywords for the other two varieties that indicate language choices that distinguish AusE and NZE from the original source variety (excluding unique regional features such as placenames and names for local flora and fauna, etc.). The resulting set includes items ranging from spelling variations, lexical choices where there are regional variants available (including ones from languages indigenous to the region), and grammatical points of difference such as choice of pronouns and modal verbs. The selected variants from the BrE standard have then been traced diachronically (using the complete Hansard record for the period) to assess the origin of the change, and track the direction of influence between Australia and New Zealand. These points of difference have, in turn, been investigated in a corpus of current (2015) Papua New Guinea (PNG) Hansard, to assess the extent of potential epicentral influence from the country's geographically closest English-speaking neighbours, as opposed to the residual influence from BrE. Results indicate some clear local norms developing in both AusE and NZE, with the direction of influence often clearly traceable via the rich diachronic resources available. The pattern of influence is less clearcut in the PNG data, suggesting the need for a larger dataset to explore epicentral effects for this, and possibly other emerging varieties of English in the region, such as those in Fiji and the Solomon Islands.

Reference

Kruger, H., van Rooy, B. & Smith, A. 2019. Register change in the British and Australian Hansard (1901-2015), *Journal of English Linguistics* 47(3): 183-220

La Peruta, Roberta

BE THAT AS IT MAY: Epicenters of cross-border dis/similarities in the use of the subjunctive In Canadian vs. American English

Recent studies focusing on the distribution of the mandative subjunctive (henceforth MS) in Present Day English show that its frequency of use has been increasing in formal written American English (AmE) from the last century onwards (i.a. Turner 1980, Crawford 2009, Schlüter 2009). Conversely, previous research on British English (BrE) found that this variety lags behind in the use of the subjunctive, broadly favoring modal periphrastic constructions, whereas Australian English (AuE) and New Zealand English (NZE) seem to steer a middle course between AmE and BrE (e.g. Övergaard 1995; Hundt 1998). However, this alleged 20th century revival led by the Americans has not yet been properly investigated in other settler varieties, such as Canadian English (CanE). To fill this gap in the research, the present study sets out to explore the MS in CanE and AmE, seen as a potential linguistic epicenter, viz. "as a model of English for (neighbouring?) countries" (Hundt 2013: 185). Based on a quantitative analysis of the Strathy Corpus of Canadian English (Strathy) and the Corpus of Historical American English (COHA), this paper diachronically tracks the trajectory of use of the MS across the US-Canada border, providing evidence from a cross-varietal sociolinguistic standpoint, and taking into consideration language external-factors such as year, genre, historic context, political history, local language policies, social influence, and economic relations between Canada and (arguably) the most influential L1 country, that clearly affect ongoing linguistic variation. The relevance of this contribution primarily stems from the fact that no previous research has so far focused on a contrastive and diachronic comparison between CanE and its alleged epicenter in this regard. The collection of evidence from big data, along with the systematic inclusion of language-external variables, gathers insights and pinpoints fruitful suggestions regarding American transnational epicentral influence on its neighbor, as well as cross- border dis/similarities concerning the subject under analysis. Furthermore, the size of the selected data allows for a robust analysis that hints at important implications on another hot (linguistic) debate, the so-called process of ongoing "Americanization" in CanE (Boberg 2004). Following a Labovian variationist approach to language, this paper allows to take an original snapshot in the current situation of the world's most pluricentric language and to study the sociolinguistic *milieu* of the varieties mentioned, while offering critical insights on the current discussion on linguistic epicenters in the World Englishes paradigm.

References

Boberg, C. (2004). English in Canada: phonology. In Kortmann, Bernd, Edgard W. Schneider, Kate Burridge, Rajend Mesthrie and Clive Upton, eds. *A Handbook of Varieties of English*. Volume 1: Phonology, 351-365. Berlin: de Gruyter.

Crawford, W. J. (2009). The mandative subjunctive. In G. Rohdenburg and J. Schlüter (ed.) *One Language, Two Grammars?: Differences Between British and American English*, 257–276. Cambridge: Cambridge University Press.

Davies, M. (2013). *Corpus of Global Web-Based English: 1.9 billion words from speakers in 20 countries*. Available online at http://corpus.byu.edu/glowbe/.

Hundt, M. (1998). It is important that this study (should) be based on the analysis of parallel corpora: on the use of the mandative subjunctive in four major varieties of English in H. Lindquist (ed.) *The Major Varieties of English: papers from Maven* 97(1), 159–175 Växjö: Acta Wexionensia.

Hundt, M. (2009). Colonial lag, colonial innovation, or simply language change? In G. Rohdenburg and J. Schlüter, (ed.), *One Language, Two Grammars?: Differences Between British and American English,* 13–37. Cambridge: Cambridge University Press.

Hundt, M. (2013). The diversification of English: Old, new and emerging epicentres. In Daniel Schreier and Marianne Hundt (eds.), *English as a Contact Language*, 182-203. Cambridge: Cambridge University Press.

Hundt, M. (2018). It is time that this (should) be studied across a broader range of Englishes:

a global trip around mandative subjunctives. In Sandra C. Deshors, Modeling *World Englishes. Assessing the interplay of emancipation and globalization of ESL varieties*, 217-244. Amsterdam: John Benjamins Publishing.

Övergaard G. (1995). The Mandative Subjunctive in American and British English in the 20th Century Stockholm: Almqvist & Wiksell.

Peters, P. (1998). The survival of the subjunctive: Evidence of its use in Australia and elsewhere.

English World-Wide: A Journal of Varieties of English 19(1), 87–103.

Schlüter, J. (2009). The conditional subjunctive. In G. Rohdenburg and J. Schlüter (ed.), *One Language, Two Grammars?: Differences Between British and American English,* 277–305. Cambridge: Cambridge University Press.

Serpollet, N. (2001). The mandative subjunctive in British English seems to be alive and kicking... Is this due to the influence of American English? In *Proceedings of the Corpus Linguistics 2001 Conference,* ed. P. Rayson, A. Wilson, T. McEnery, A. Hardie.

Turner, J. F. (1980). The marked subjunctive in contemporary English. *Studia Neophilologica* 52: 271–277

Mair, Christian (Freiburg)

Global Jamaican: Migration, media and the creation of an informal epicentre

From the 18th century to the end of the colonial period, Jamaica was the demographic and cultural centre of the British West Indies. During this period, the island played a considerable role in the spread of Creole and English throughout the wider Caribbean. These early encounters have left a rich legacy in the vernacular linguistic landscapes of the region, but – as I will show on the basis of the OED – the Jamaican input into English beyond the Caribbean remained marginal.

This has profoundly changed since the end of World War II. Successive waves of late colonial and postcolonial migrations led to the establishment of large diasporic communities in the UK, the US and Canada. Jamaican Creole in particular, rather than wither away, transformed the sociolinguistic landscape of the immigrant destinations. The rich literature describes new hybrid vernaculars such as 'London Jamaican' (Sebba 1993) or 'British Creole' (Patrick 2004, Sebba 2004). Jamaican Creole figures prominently in the literature on sociolinguistic 'crossing' (Rampton 1995), and there is significant Jamaican input in contemporary urban multiethnolects, not only in the Global North (see Cheshire et al. 2011 on Multicultural London English and Kerswill 2014 on the construction of 'Jafaican' in folk linguistics), but also in African megacities such as Lagos, Nairobi and Johannesburg. The paper will document the extent and types of Jamaican influences on World Englishes on the basis of GloWbE and other web-sourced corpora and then discuss the role of migration, globally popular subcultures of Jamaican origin (e.g. Rasta and reggae) and the World-Wide Web/social media in general as driving forces of language contact.

Finally, I will draw on Mair's (2013) four-tier stratification of varieties of English into 'hyper-central', 'super-central', 'central' and 'peripheral' ones, to explore the question how well this model might accommodate the 'epicentral' effects discussed here.

References:

Cheshire, Jenny, Paul Kerswill, Sue Fox, and Eívind Torgersen. 2011. "Contact, the feature pool and the speech community: the emergence of Multicultural London English." *Journal of Sociolinguistics* 15: 151-196.

Kerswill, Paul. 2014. "The objectification of 'Jafaican': the discoursal embedding of multicultural London English in the British media." In Jannis Androutsopoulos, ed. *Mediatization and sociolinguistic change*. Berlin: de Gruyter. 427-456.

Mair, Christian. 2013. "The World System of Englishes: accounting for the transnational importance of mobile and mediated vernaculars." *English World-Wide* 34: 253-278.

Patrick, Peter L. 2004. "British Creole: phonology." In Bernd Kortmann and Edgar Schneider, eds. *Handbook of varieties of English.* Vol. 1. Berlin: Mouton de Gruyter. 231-243.

Rampton, Ben. 1995. *Crossing: language and ethnicity among adolescents*. London: Longman.

Sebba, Mark. 1993. *London Jamaican*. London: Longman.

Sebba, Mark. 2004. "British Creole: morphology and syntax." In Bernd Kortmann and Edgar Schneider, eds. *Handbook of varieties of English*. Vol. 2. Berlin: Mouton de Gruyter. 196-208.

Meierkord, Christiane (Ruhr-University of Bochum)

Still in awe of the colonial variety? Post-protectorate Uganda as a challenge to current models of the influence between Englishes

More often than not, models of world Englishes have been developed on the basis of observations made for post-colonial varieties of English. This is the case with Schneider (2007), Mair (2013) and also with epicentre theory (see Peters 2009 and Gries & Bernaisch 2016). However, varieties used in contexts that escape the traditional ENL-ESL-EFT distinction, such as those used in post-protectorates, seem to pose challenges to these models.

This paper presents the overall results of a research project that has investigated selected features of Ugandan English. Based on 90 face-to-face conversations, which are now part of ICE-Uganda, it has examined how speakers in the cities of Gulu (having a Nilotic first language), Kampala and Mbarara (with Bantu first languages) express futurity, ability and obligation.

At the risk of overgeneralisation, findings indicate that speakers overwhelmingly prefer *will* as a marker of futurity, whilst *going to* is used considerably less. As regards ability, there is a strong preference for *can*, but Bantu speakers also use *could* in 12% of all cases, whilst this is only used in 5% by Nilotic speakers in Gulu. To express obligation, *have to* is the most frequent expression, with 57/58% in Mbarara and Kampala but 46% in Gulu, where *should* is preferred in 30%.

Forms that the literature has identified in the 'hub' variety American English or the hypercentral varieties Indian or Nigerian English, and whose use would indicate influence from these, such as *might* to express futurity (reported for Nigerian English), *could* for ability or *may* for obligation (reported for Indian English, see Mesthrie & Bhatt 2008) are used seldomly. Neither is there an overuse of *shall* or a distinctive use of *would* instead of *will* to express futurity, as reported for Kenyan English in Buregeya (2019).

The lack of such influence challenges the assumption of epicentre theory, according to which an endonormatively stable variety influences varieties in neighbouring countries, as well as Mair's (2013) assumption that varieties that have a high impact in terms of the demographic and economic weight of their speakers influence those with a lower impact. Rather, exonormative orientation towards British English, due to the predominantly formal acquisition of English in Uganda, and first

language influence seem to determine how many variants are tolerated and which ones are preferred.

There is, on the other hand, variation regarding the place of recording, i.e. Gulu in the Nilotic speaking North of Uganda, and Kampala and Mbarara in the Bantu- speaking Central and Western regions. Gulu speakers, with their Nilotic languages having only one expression to formally mark futurity, have significantly higher shares of *will*, whilst *going to* is used more frequently in Kampala. Similarly, Nilotic speakers recorded in Kampala and Mbarara have considerably higher uses of *have to* (66%).

The observed influence of the Bantu speakers' behaviour on that of the Nilotic speakers residing in Bantu-speaking areas confirms Meierkord's (2012) finding for South Africa, that regular long-term exposure to another variety results in accommodation, making this a factor that needs to be better accounted for.

References

Buregeya, Alfred. 2019. *Kenyan English*. Boston & Berlin: Walter de Gruyter.

Gries, Stefan Th. & Bernaisch, Tobias. 2016. ‘Exploring epicentres empirically: focus on South Asian Englishes’, *English World-Wide* 37(1): 1–25.

Meierkord, Christiane. 2012. *Interactions across Englishes. Linguistic Choices in Local and International Contact Situations.* Cambridge: Cambridge University Press.

Mair, Christian. 2013. ‘The world system of Englishes: accounting for the transnational importance of mobile and mediated vernaculars', *World Englishes* 34(3): 253–278.

Mesthrie, Rajend and Bhatt, Rakesh. 2008. *World Englishes. The Study of New Linguistic Varieties*. Cambridge: Cambridge University Press.

Peters, Pam. 2009. 'Australian English as a regional epicentre'. In Thomas Hoffman & Lucia Siebers (eds.), *World Englishes—Problems, Properties and Prospects.* Amsterdam: John Benjamins, pp. 107–124.

Peters, Pam, Adam Smith and Minna Korhonen

"A diachronic study of *-ed/-t* inflectional variation in Australian and New Zealand English: Epicentral influence under the radar?"

Variation in the past tense/participle ending *–ed/-t* is associated with a small set of verbs: monosyllabic verbs ending in *l, m, n, p*, including *spell, dream, burn, leap.* In American English these are usually inflected with *–ed*, whereas in British English there is greater variation with *–t* (Levin 2009). Meanwhile C21 data from GloWbE for Australian and New Zealand English show even more extensive of the *–t* forms than in the two international forms of English*,* at least in the case of *burn* and *earn* (Peters & Burridge 2020). In other previous studies based on the ICE corpora, the level of variation (use of the *–t* form) was higher in Australian than New Zealand, except for the lowest frequency verbs (Peters 2009). How far back in time these differences go in the antipodes not known (cf Anderwald 2016), nor whether they apply to the whole candidate set of 12 verbs. A further question to be investigated is whether the *–ed/-t* spellings tend be associated with the past tense and past participle respectively. Diachronic evidence is needed to examine these questions, which would contribute to wider discussion of the epicentral relationship between Australian and New Zealand English in their pre-endonormative phases (cf Hundt 2013). The data will also be used to address issues such as how far the degree of *–ed/-t* variability correlates with the relative frequency of each verb in the two varieties in C21 (Bybee 1995).

In this study we compare the frequencies in GloWbE of the forms of the past tense and past participles used for all 12 verbs in C21 Australian and New Zealand English, with those in other corpus data from C19 and C20. For C19 Australian English we draw on the multigeneric corpus COOEE (Fritz 2010), and for C19 New Zealand English, its Hansard record commencing in 1854. For C20 English of both varieties we use data from their respective Hansards.

Preliminary data from COOEE suggest that the extended use of *-t* spellings originated in C19 Australian English with the higher frequency examples (*burn, dream)*, independent of their increased use in later C20 British English noted by Gowers (1965). Greater use of *–t* in Australian English in both centuries is to be expected, given its higher rates of usage in GloWbE data from C21, in comparison with New Zealand English. Using diachronic data from COOEE and Hansard we will be able to compare the patterns of *–ed/-t* usage for the two past forms of the 12 verbs in both varieties during C19 and C20. It is of no small interest to see whether the use of *–t* has been consistently higher in Australian English since C19, and capable of influencing New Zealand English usage over a short- or longer period.

The GloWbE data showing that *–t* usage is more common now in Australian than New Zealand English gives grounds for associating epicentral influence with this element of verb morphology. Research on contemporary usage of shared lexical innovations in neighboring varieties of English has shown how they maintain higher levels of usage in the variety where they originated, which points to the likely direction of influence within a notional epicentre (Peters et al. 2019). This morphological study would add further evidence to show that a regional variety can exercise epicentral influence on a neighboring variety before attaining its endonormativity.

References

Anderwald, L 2016 *Language Between Description and Prescription. Verbs and Verb Categories in Nineteenth-Century Grammars of English (Oxford Studies in the History of English 6* Oxford University Press

Bybee, J 1995. Regular morphology and the lexicon*. Language and Cognitive Processes* 10, 425–455

Fritz, C 2010 A short history of Australian spelling *Australian Journal of Linguistics* 30:2 *227-182*

Gowers, E 1965 (ed) *Fowler's Dictionary of Modern English Usage* Oxford, Clarendon Press

Hundt, M 2013 The diversification of English: Old, new and emerging epicentres. In D Schreier & M Hundt (eds) *English as a Contact Language.* 182-203 Cambridge, Cambridge University Press

Levin, M 2009 The formation of the preterite and the past participle. In Rohdenburg and Schluter eds *One Language, Two Grammars* Cambridge, Cambridge University Press

Peters, P 2009 Irregular verbs: regularisation and ongoing variability. In P Peters, P Collins & A Smith (eds) *Comparative Studies of Australian and New Zealand English: Grammar and Beyond* Amsterdam, John Benjamins.

Peters, P & Burridge, K 2020 English in Australia -- Extraterritorial influences. In S Buschfeld & A Kautzsch (eds) *Modelling World Englishes: A joint approach to Postcolonial and non-Postcolonial Englishes* Edinburgh, Edinburgh University Press

Peters, P, Smith, A & Bernaisch, T 2019 Shared lexical Innovations in Australian and New Zealand English *Dictionaries* 40(2) 1-25

Schneider, Edgar W. (Universität Regensburg)

Parameters of epicentral status: requirements and traces

The notion of linguistic "epicenters" derives from the concept of "pluricentric" languages, introduced by Kloss (1978) and popularized by Clyne (1992) and, in English linguistics, Leitner (1992). It has been used somewhat fuzzily, so I argue there are two distinct readings, a weak one (claiming that several national varieties of languages co-exist, with norms of their own) and a strong one (assuming that some language varieties exert influence on others, typically in their vicinity). It is intuitively appealing but appears questionable at second glance, and, most importantly, empirically difficult to pin down. Extrapolating from significant work by Hundt (2013) and others, the present paper surveys and discusses several parameters that define the concept, particularly its strong reading, thus offering steps towards a theory of pluricentricity and epicenters. Whenever possible, the factors will be exemplified using data from cases in point which have been suggested, based on the literature and on selected corpus data. I believe it is important to distinguish three different types of relevant perspectives: (a) defining factors, i.e. prerequisites of epicentral influence; (b) methodological issues, i.e. problems of identifying possible influences; and (c) further issues, i.e. pertinent questions to be asked.

Defining parameters include the following:

- Size and demography: Larger nations influence smaller ones (e.g. Australia influences New Zealand; Peters 2009).
- Proximity: Influence is strongest on nearby locations and nations (e.g. of India on Sri Lanka).
- Developmental and economic status: more developed and wealthier nations influence less developed ones (e.g. Singapore in South-East Asia).
- Intensity of exchange: The more institutionalized interaction is the higher is the probability of influence.
- Monodirectionality: It is assumed that influences proceed only in one direction, but don't contact settings always produce mutual impact?
- Awareness and attitudes: It is not clear whether speakers are aware of (or need to be aware of) one variety exerting influence upon another. Based on known parameters of language contact it seems likely that conscious awareness is not required but some sort of a positive attitude is supportive and perhaps necessary for components of another culture and variety to be mirrored.

Methodological issues of identifying traces of epicentral impact are:

- Synchronicity: Are comparisons of modern data (e.g. ICE-corpora) sufficient to posit epicentral influence, or would diachronic evidence be mandatory?
- Nature of evidence: To posit monodirectional linguistic influence, is it sufficient to observe the mere presence of the same features, are similar feature frequencies required, or are similar constraint effects most convincing?
- Practical issues and connections: Can epicentral influences be substantiated by finding explicit contacts (such as the American Thomasites coming to the Philippines, or Indian textbooks being employed in Sri Lanka)?

Additional aspects that can be considered meaningfully are:

- Standard-ness: Can only standard varieties serve as epicenters (which would imply that epicenters should have reached the stage of endonormative stabilization, in itself a multifaceted concept), or are nonstandard influences also effective (as illustrated by features of African American English used in Japanese Hip Hop)?
- Perception or production? The notion of pluricentricity is briefly compared to "pluriareality", a concept hotly debated in German linguistics (Dollinger 2019), and I argue that while

pluriareality focuses on production pluricentricity is rooted more strongly in the perception of variety differences.

Ultimately, the above considerations strongly suggest that epicentral influence is to be seen not as an all-or-nothing effect but as a "prototypical" concept, a relationship which may hold to a lesser or stronger extent and which in turn is composed of and can be detected by a range of composite factors.

References

Clyne, Michael, ed. 1992. *Pluricentric Languages. Differing Norms in Different Nations*. Berlin, New York: Mouton de Gruyter.

Dollinger, Stefan. 2019. *The Pluricentricity Debate: On Austrian German and Other Germanic Standard Varieties*. New York, London: Routledge.

Hundt, Marianne. 2013. "The diversification of English: Old, new and emerging epicentres." In Daniel Schreier & Marianne Hundt, eds., *English as a Contact Language*. Cambridge: Cambridge University Press, 182-203.

Kloss, Heinz. 1978. *Die Entwicklung neuer germanischer Kultursprachen seit 1800*. 2nd ed. Düsseldorf: Schwann.

Leitner, Gerhard. 1992. "English as a pluricentric language." In Clyne, ed. 1992: 179-237.

Peters, Pam. 2009. "Australian English as a regional epicentre." In Thomas Hoffmann & Lucia Siebers, eds. *World Englishes – Problems, Properties and Prospects*. Amsterdam: John Benjamins, 107-12.

Workshop 5: To boldly go: Corpus approaches to the language of Science Fiction

Convenors: Claudia Lange (TU Dresden) Sofia Rüdiger (University of Bayreuth)

Engaging with possible futures is an essential human endeavor and the popularity of the Science Fiction (SF) genre in general, but also particularly among linguists, thus does not come as a surprise. SF has the power to continually shape, stimulate, and challenge contemporary thought and societal norms, and serves much deeper undertakings than being mere speculative fiction.

While being notoriously difficult to define, most writers on Science Fiction (SF) take Darko Suvin's by now classic definition of the genre as a reference point (e.g. Shippey 2007: 15, Adams 2017: 331):

SF is, then, a literary genre whose necessary and sufficient conditions are the presence and interaction of estrangement and cognition, and whose main formal device is an imaginative framework alternative to the author's empirical environment. (Suvin 2016[1979]: 20)

Science Fiction texts – which we take to include stories, novels, fan fiction, video games, TV series, and movies – rely on linguistic "means of estrangement" listed by Adams (2017: 333ff.) to different degrees. Science Fiction's alternative 'imaginative framework' comes to life via the creative use of language and may range from occasional 'alien' referring expressions to the development of fully-fledged artificial languages, with Klingon being the most iconic and enduring example (Adams 2011, Okrent 2009).

Whereas previous research on SF has rested mainly on literary and qualitative approaches, we propose corpus linguistics as a fruitful method to investigate the language of Science Fiction from new perspectives. Recently published resources, such as the BYU TV Corpus (2019) and the BYU Movie Corpus (2019), include ample SF material, offer easy access to telecinematic discourse, and have yet to be employed for large-scale corpus linguistic research of SF.

References

Adams, Michael (2017), The pragmatics of estrangement in fantasy and science fiction. In Locher, Miriam A. & Andreas H. Jucker (eds.), *Pragmatics of Fiction*. Berlin: De Gruyter Mouton, 329-63.

Adams, Michael (ed.) (2011), *From Elvish to Klingon: Exploring Invented Languages*. Oxford: Oxford University Press.

Bould, Mark (2009), Language and Linguistics. In Bould, Mark, Butler, Andrew M., Roberts, Adam & Vint, Sherryl (eds.), *The Routledge Companion to Science Fiction*. London: Routledge, 225-235.

Epstein, Robert, Roberts, Gary, & Beber, Grace (eds.) (2008); *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*. Dordrecht: Springer.

Okrent, Arika (2009), *In the Land of Invented Languages: Esperanto Rock Stars, Klingon Poets, Loglan Lovers, and the Mad Dreamers Who Tried to Build a Perfect Language*. New York: Spiegel & Grau.

Peterson, David J. (2015), *The Art of Language Invention: From Horse-Lords to Dark Elves, the Words behind World-Building*. New York: Penguin.

Shippey, Tom (2007), Hard Reading: The Challenges of Science Fiction. In Seed, David (ed.), *A Companion to Science Fiction*. Malden, Oxford: Blackwell Publishing Ltd., 9-26.

Suvin, Darko (2016 [1979]), *Metamorphoses of Science Fiction: On the Poetics and History of a Literary Genre*. Oxford: Peter Lang.

The Movie Corpus (2019), https://www.english-corpora.org/movies/

The TV Corpus (2019), https://www.english-corpora.org/tv/

Behrens, Tanja & Dr. Anke Schulz, Universität Bremen, Fachbereich 10 – Sprach- und Literaturwissenschaften

Where no woman has gone before – five decades on the *Enterprise*

The role of women in the *Star Trek* series has been investigated from several perspective (Blair 1983, Joyrich 1996, Kubitza 2016), but not from a linguistic one to any satisfying extent. Thus, our study describes the use of spoken language of the women and men in the *Star Trek* TV series. We use 14 episodes of the original *Star Trek* series from 1966-7 (about 49,000 words) as our corpus and compare this to a second corpus of five episodes of the new *Star Trek Discovery* series from 2017 (about 17,000 words). More than 50 years lie between these two series, and in these five decades the feminist movement made many achievements in the real world, all around the world. These developments have changed the way women speak, and the way that women are spoken to. Is this change reflected in this particular Science Fiction TV series?

With corpus stylistics methods (McIntyre & Walker 2019), we describe the language of the women and men on board the Enterprise and the Discovery. In particular, we are looking at clause length, sentence type and interruptions, which are manually annotated with the UAM corpus tool (O'Donnell 2019) in four sub-corpora of 3,000 words each (women 1967, men 1967, women 2017 and men 2017 scripts).

In a second step, we analyze the different ways that female and male protagonists are addressed, n-grams specific to female and male protagonists and selected Key words in context (KWIC) with the AntConc software (Anthony 2019).

Without doubt, Science Fiction has the power to stimulate and challenge contemporary thought and societal norms. Our results, however, seem to suggest that the world of *Star Trek* is sometimes more archaic than it would probably want to be.

References

Anthony, L. (2019). AntConc (Version 3.5.8) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Blair, Karin. 1983. “Sex and ‘Star Trek’ ”. In: *Science Fiction Studies*, Vol. 10, No. 3, pp. 292-297.

Joyrich, Lynne. 1996. “Feminist Enterprise? ‘Star Trek: The Next Generation’ and the Occupation of Femininity”. In: *Cinema Journal*, Vol. 35, No. 2, pp. 61-84.

Kubitza, Nicole. 2016. *Pretty in Space. Die Frauendarstellung in Star Trek und anderen US-amerikanischen Dramaserien der 1960er Jahre.* Göttingen: V&R.

McIntyre, Dan and Walker, Brian. 2019. *Corpus Stylistics: Theory and Practice*. Edinburgh: Edinburgh University Press.

O'Donnell, Michael (2019). UAM corpus tool (Version 3.3) [Computer Software]. Madrid, Spain: Universidad Autónoma de Madrid. Available from http://www.corpustool.com/download.html

Blombach, Andreas, Thomas Proisl, Stefan Evert, Philipp Heinrich, Natalie Dykes

Computational Corpus Linguistics Group Bismarckstr. 6, 91054 Erlangen, Germany Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) {andreas.blombach, thomas.proisl, stefan.evert, philipp.heinrich, natalie.mary.dykes}@fau.de

Into the *Perryverse*: A CL Journey to the Realm of Lexical Complexity

Perry Rhodan is the eponymous hero of a German science fiction series which has been published continuously and uninterrupted since 1961. (There have also been uninterrupted runs of translations in France, Japan and the Netherlands, as well as more limited runs in several other countries.) Short novels are published weekly in *Heftroman* form (roughly equivalent to dime novels or pulp magazines). To date, the main series comprises over 3,000 of these novels, which justifies its own designation as the world's "biggest science fiction series". The series has also spawned several spin-offs and "proper" novels in its so-called *Perryverse*, as well as hardcover collections, audio and comic books.

In previous research, we evaluated measures for different dimensions of textual complexity (e.g. lexical diversity, lexical disparity and syntactic complexity; for the latter, see Proisl et al. 2019) by comparing i.a. German highbrow and lowbrow literature.[1] Books classified as highbrow literature (selected for having been nominated for prestigious literature prizes) could quite reliably be separated from dime novels from the crime ("Jerry Cotton"), romance ("Julia Extra") and horror ("Geisterjäger John Sinclair") genres.[2] However, Perry Rhodan novels, representing science fiction dime novels, appeared to be much more complex, especially lexically, than their counterparts from other genres (and even more complex than many highbrow novels, according to some measures).

We now want to take a closer look at this apparent complexity, focussing especially on the series' idiosyncratic vocabulary (e.g. *Raumer*, *Impulsstrahler*, *Mausbiber*, *Arkonide*, *Linearraum*, *Zellaktivator*). Is it just the sheer amount of invented technical terms, species, planets and proper names that increases measured complexity, or is there more to it? In other words, does Perry Rhodan – and possibly science fiction in general – only appear especially complex to the uninitiated, or is it similar for regular readers familiar with its peculiar language? How can we characterise and categorise this special vocabulary?

To find out to which extent our results may be generalised (e.g. whether they are language-specific or even limited to Perry Rhodan), we compare English science fiction texts with texts from other genres, using both published works and fan fiction.

We also want to track changes over time in style, complexity and vocabulary, taking into account differences between individual authors. Our German corpus currently consists of approx. 650 Perry Rhodan novels from 1961 to 1975. The novels from this period, especially the earlier ones, are known for their focus on human expansionism and war, whilst later installments more often feature diplomacy and peaceful resolutions of conflicts. We plan to investigate if we can identify such changes using techniques like semantic tagging and topic modelling.

Additionally, we want to compare the older novels to Perry Rhodan Neo, a modern retelling of the original story, which has been published biweekly since 2011, in parallel to the main series.

Reference

Proisl, Thomas, Leonard Konle, Stefan Evert, and Fotis Jannidis. 2019. "Dependenzbasierte syntaktische Komplexitätsmaße." In *DHd 2019 Digital Humanities: multimedial & multimodal. Konferenzabstracts*, edited by Patrick Sahle, 270–273. https://doi.org/10.5281/zenodo.2596095

[1] Using our toolbox available at https://github.com/tsproisl/Linguistic_and_Stylistic_Complexity.

[2] We want to thank our colleagues and collaborators at the University of Würzburg for compiling this corpus: Fotis Jannidis, Leonard Konle and Steffen Pielström.

Eberl, Martin (LMU München)

How “alien” are fictional languages? Approaching ‘estrangement’ in constructed languages

One of the pillars of science fiction is the encounter with the unknown, such as alien beings and foreign cultures, and the pursuit of the question of what else might be out there. For a linguist, it is difficult to suspend disbelief when the newly encountered beings perfectly speak the protagonists' language, or to avoid disappointment when the language barrier is quickly overcome by a universal translator system or magic spell. Even when included in detail, many constructed languages (“conlangs”) in sci-fi and fantasy bear striking similarity to existing natural languages, keeping, for example, the basic aspects of English lexis and morphology while merely changing prototypical word order. Instances of truly foreign language systems such as in *Arrival* are a rare exception.

This is surprising, since, as J.R.R. Tolkien put it in a 1931 lecture, in creating a language from scratch, “[there are] no base considerations of the 'practical', the easiest for the 'modern mind', or for the million, only a question of taste, a satisfaction of a personal pleasure”. What then, could constrain language creators, and how?

This paper explores the influence of one’s native language as a possible factor. A corpus of material from 50 constructed languages – 30 by English L1 speakers, and 10 each by French and German L1 speakers – was built. It encompasses both primary source material and grammatical descriptions provided for these languages and includes both languages created by linguists and hobbyists as well as both those known to a large audience through literature or television and those created by hobbyists purely for their own pleasure. Each conlang was then scrutinized in regard to various features in the areas of phonology, orthography, morphology, syntax and semantics.

Patterns in a conlang’s deviation from its’ creator’s L1 reveal that the influence is diverse: first, traditional linguistic features such as phonology or morphology are more often affected on a systemic rather than a feature level, and some features seem entirely free of native language influence. Second, there are strong parallels between L2 acquisition and language creation. Third, creators often retain features they perceive as common across languages, whereas they tend to adjust or delete features they deem ‘characteristic’ of their L1. Last, conceptual thinking patterns such as numeral bases or spatiotemporal features of the L1 are often mirrored in the conlang.

References:

Adams, Michael (ed.) (2011), From Elvish to Klingon: Exploring Invented Languages. Oxford: Oxford University Press.

Couturat, Louis and Léopold Leau. 1903. Histoire de la langue universelle. Paris, Hachette.

Okrent, Arika (2009), In the Land of Invented Languages: Esperanto Rock Stars, Klingon Poets, Loglan Lovers, and the Mad Dreamers Who Tried to Build a Perfect Language. New York: Spiegel & Grau.

Peterson, David J. (2015), The Art of Language Invention: From Horse-Lords to Dark Elves, the Words behind World-Building. New York: Penguin.

Tolkien, John Ronald Reuel. 1931 (1983). “A Secret Vice”. The Monsters and the Critics and Other Essays. Ed. Christopher Tolkien. London: George Allen & Unwin. 198-223

Gee, Matt (Birmingham City University)

“there was much new to grok”: A semantic analysis of word creation in science fiction.

As can be witnessed in projects such as *The Oxford Dictionary of Science Fiction* (Prucher 2007), science fiction has been fertile ground for the creation of new words and concepts. Whereas the aforementioned dictionary was constructed by eliciting examples and citations from volunteers, this paper presents an initial foray into corpus-based methods for uncovering lexis unique to science fiction. In addition, this study seeks to examine how such neologisms are formed and the means by which authors impart their meaning to the reader, drawing on frameworks of semantic word relations and work in Cognitive Linguistics (e.g. Rosch et al. 1976) in the process.

Three stages of analysis are presented. The first is based on a corpus of science fiction texts scanned from physical copies of novels and short-story collections. The corpus consists of 3,107,299 tokens covering the period 1890-1994 and is POS-tagged using the Stanford CoreNLP tagger. Words unique to the science fiction texts are extracted by comparing the corpus against the British National Corpus (BNC Consortium 2007). The words are grouped into grammatical categories which shows that nouns make up the majority of neologisms (254 types), with new adjectives (42) and verbs (11) being less frequent. However, it is noted that the POS-tagging produces mixed results with science fiction language for various reasons.

The second stage of analysis investigates the morphological rules which underpin the formation of the neologisms, including compounding (e.g. *bloodmother*, *copseyes*), derivation (e.g. *hypersleep*) and lexical blending (e.g. *pedecab*). However, not all the discovered words have distinguishable morphological roots and, even when they do, it may not be possible to determine their meaning without further context.

The third stage investigates the occurrences of each word in context. This reveals the use of definitions and glosses by the authors, both in narration and direct speech. Where a definition is not provided, the reader is left to determine the meaning of the word from the context. Some suggestions are made as to the kinds of clues available to the reader, including co-occurrence with synonyms or antonyms (e.g. *precog* co-occurs with semantically related words) and the drip-feeding of attributes pertaining to the concept being referenced (e.g. the physical properties of *triffids* being revealed over several sentences). In addition, authors may deliberately keep the meaning of a word vague for large portions a novel (as shown with *shifgrethor* and *grok*), supporting themes of alienation and otherworldly-ness.

References

BNC Consortium. 2007. *The British National Corpus, version 3 (BNC XML Edition)*. Distributed by Bodleian Libraries, University of Oxford. http://www.natcorp.ox.ac.uk.

Prucher, Jeff. 2007. *Brave New Words: The Oxford Dictionary of Science Fiction*. Oxford: Oxford University Press.

Rosch, Eleanor, Caroline Mervis, Wayne Gray, David Johnson and Penny Boyes-Braem. 1976. “Basic objects in natural categories.” *Cognitive Psychology* 8: 382-439

Laliberté, Catherine, Melanie Keller and Diana Wengler (LMU Munich)

"So I trucked out to the border, learned to say ain't*, came to find work": the sociolinguistics of* Firefly

Firefly is a science-fiction TV series created by Joss Whedon, which originally aired in 2002 and 2003 in the United States. Like *Star Wars* and *Cowboy Bebop*, *Firefly* belongs specifically to the space western sub-genre, which allies traditional science fiction and Western tropes and aesthetics by layering, in this case, a dystopian society, space travel, stand-offs in desolate landscapes, train robberies and saloon brawls. This juxtaposition of genres is reflected in the linguistic behavior of *Firefly*'s characters in three major ways: world-specific slang and jargon, Chinese code-switching (see Mandala 2010), and features mirroring Southern American English. This study explores the latter employing the methods of variationist sociolinguistics (see Tagliamonte 2012).

Using a corpus of all fourteen episodes of the series as well as the feature film *Serenity* (2005), we show that the frequency of features reminiscent of non-standard (Southern) varieties, such as nasal fronting (see Thomas 2008), negative concord and BE- and DO-leveling (see Murray & Simon 2008), not only correspond to "real-world" constraints but are also stratified according to the social realities of the world of *Firefly*, with rebel smugglers exhibiting higher rates of variants perceived as non-standard than characters representing valued professions.

This pattern contributes subtly to world-building in that it indexes divisions between the developed, superpower-controlled territories (the Core) and the wilder, recently-settled edge of the universe (the Rim). The use of Southern features for speech associated with the Rim evokes the Wild West, and at the same time draws on present-day notions of linguistic (non-)standardness to indicate social and physical marginality. *Firefly* therefore relies on its audience's pre-existing linguistic knowledge of the real world to create its fictional one.

References

Mandala, Susan. 2010. *Language in Science Fiction and Fantasy: the question of style*. New York: Continuum.

Murray, Thomas E., and Beth Lee Simon. 2008. Colloquial American English: grammatical features. In Edgar W. Schneider, ed. *Varieties of English 2: The Americas and the Caribbean*. Berlin/New York: De Gruyter, 401-27.

Tagliamonte, Sali A. 2012. *Variationist sociolinguistics: change, observation, interpretation*. Malden: Wiley-Blackwell.

Thomas, Eric R. 2008. Rural Southern white accents. In Edgar W. Schneider, ed. *Varieties of English 2: The Americas and the Caribbean*. Berlin/New York: De Gruyter, 87-114.

Oakes, Elizabeth (University of Helsinki)

Representing Altered Consciousness in American New Wave Science Fiction: A Corpus-Stylistic Analysis

During the 1960s and 1970s, style became an increasingly central concern in science fiction as authors, editors, and academics began to perceive and promote the genre as literary art (Broderick, Mandala 3-7). Concurrently, the depiction of altered states of consciousness, such as dreams, drug trips, and madness, increased in science fiction. No longer merely a device to plug a plot hole or vilify the villain, altered consciousness became thematically central to exploring sociological, anthropological, and psychological issues.

As part of a larger project studying representations of altered consciousness in American New Wave science fiction, this paper explores the language depicting altered consciousness in such novels published between 1960 and 1964 using a corpus-stylistic approach. Analysis with dictionary programs that categorize the words of a text according to rhetorical and psychological paradigms reveals lexical aspects of style over a corpus of novels assembled from Hugo and Nebula winning authors of the period, such as Ursula K. Le Guin and Philip K. Dick. This paper describes similarities in representing altered states of consciousness across the corpus.

After mapping categorical lexical tendencies across the entire corpus, traditional stylistic analysis describes them in conjunction with grammatical and syntactic aspects of style in select texts. This analysis identifies variables that distinguish different techniques of representing altered consciousness and enables construction of a model for cluster analysis on the entire corpus (see, e.g., Hope and Witmore for a similar approach). After the model has clustered texts, each group is analyzed separately with the dictionary programs to create a clear picture of the lexical tendencies of each style of representing altered consciousness.

The corpus-stylistic analysis identifies three broad strategies (for discussion of similar tendencies at the semantic level see Felman 87, 83-84, 104).

The first strategy is exaggeration or copiousness found, for example, in Ursula K. Le Guin, who packs text representing altered consciousness with higher levels of particular semantic categories, such as sensory description words.

The second strategy is reversal whereby altered states are depicted through language with opposite valence to the language portraying default states of consciousness.

The third strategy is equivalence, apparent in Philip K. Dick novels where the default state is amplified through techniques of exaggeration in order to match the style of the altered state and blend with it.

The paper concludes with a brief consideration of nascent social changes in American society during the early 1960s and science fiction’s engagement with these issues through the lens of altered consciousness.

References

Broderick, Damien. "New Wave and Backwash: 1960-1980." *The Cambridge Companion to Science Fiction.* Edited by Edward James. Cambridge University Press, Cambridge, 2003. *ProQuest*, https://search-proquest-com.libproxy.helsinki.fi/docview/2137995003?accountid=11365.

Felman, Shoshana. *Writing and Madness (Literature/Philosophy/Psychoanalysis)*. Stanford University Press, 2003.

Hope, Jonathan and Michael Witmore. "The Hundredth Psalm to the Tune of 'Greensleeves': Digital Approaches to Shakespeare's Language of Genre." *Shakespeare Quarterly*, vol. 61, no. 3, 2010, pp. 357-390. doi:10.1353/shq.2010.0002.

Mandala, Susan. *The Language in Science Fiction and Fantasy*. Bloomsbury Publishing, 2010

Ronan, Patricia (TU Dortmund) and Gerold Schneider (University of Zürich)

To boldly go – whence and where?

To boldly go where no man has gone before, popularized by the science fiction series Star Trek, has provided an iconic example for the use of split infinitives, as has also been noted by other linguists (Desagulier 2017). From its introduction, the series has arguably paved the way for the broader use of split infinitives in contemporary, informal English in spite of the structure prescribed against by grammars. The current qualitative and quantitative study aims to trace genre influence in the use of split infinitives in sci-fi and television discourses. It asks whether the use of the split infinitive in informal genres can be correlated with its prominent use in the Star Trek series.

The use of the structure, already in the introduction to first season 1966, can be seen as iconic for this series, which has reached cult-status from the late decades of the 20th century onwards. These are good conditions for a concomitant spread of a feature, first by fans of a series in in-group conversations, and later outside forums as argued by Adams (2014). However, the split infinitive is by no means a new introduction into the English language. Split infinitives have a long history in the English language and can already be found in the Middle English period (Calle-Martin 2015). However, there has been a strong prescriptivist tradition against them (Perales-Escudero 2010), which is now easing, and in contemporary language use, specific collocations, especially *to better understand*, are found in academic discourse in particular (Perales-Escudero, loc. cit.).

The current study focuses on the question how the distribution of split infinitives changes in different genres. For this, the Corpus of Historical American English (COHA), of Contemporary American English (COCA), the British National Corpus (BNC), the SOAP corpus and the Movie corpus (MOVIES) are investigated in the Brigham Young Corpus suite. In these corpora, the search strings *to* + adverb + INF are searched for and relative frequencies are observed.

The results show the influence of genre on the use of split infinitives in contemporary and near contemporary American and in British English.

References

Adams, M. 2014. Slang in new media. A case study. In J. Coleman (ed.), *Global English slang. Methodologies and perspectives*. London: Routledge. 175-186.

Calle-Martin, J. 2015. The Split Infinitive in Middle English. *NOWELE* 68.2015, Issue 2, 227 – 250.

Desagulier, G. 2017. The split infinitive: the final frontier. In: *Around the word.* https://corpling.hypotheses.org/30 (last accessed 14.10.2019)

Perales-Escudero, M.D. 2010. To Split or to Not Split: The Split Infinitive Past and Present. *Journal of English Linguistics* 39(4) 313–334

Yurchenko, Asya (Dresden University of Technology)

Quantifying the language of estrangement in science fiction using unique lexical compounds as a measure

In his paper on the pragmatics of estrangement in fantasy and science fiction, Michael Adams names the creation of new words or lexical compounds as one of the linguistic strategies for establishing estrangement (2017: 337). Some well-known examples of this are such word compounds as *warp speed* and *lightsaber*, from *Star Trek* and *Star Wars*, respectively.

An initial keyness exploration into a 250,000-token corpus consisting of episode scripts from the popular science fiction TV series *Futurama* revealed a plethora of similarly inventive and estranging noun phrases, such as *free will unit*, *mobile oppression palace*, *robot heritage*, *disembodied software*, *holographic brain*, and *flabbo-dynamic spandex*, to name a few.

Using Adams' proposal of regarding such unique lexical compounds as one of the markers of estrangement as a point of departure, the present study aims to devise a way to quantify the lexical compound inventiveness of a corpus by means of a "lexical compound (LC) estrangement score" as an extension of keyness. The study will set out to answer the following questions:

- What are the best parameters to choose when analyzing multi-word expressions for keyness and how should the initial list be narrowed down further?
- Which further statistical measures are best suited for producing the most accurate LC estrangement score?
- How accurate is the LC estrangement score when applied to other texts or corpora?

Apart from presenting and evaluating a methodology for the quantification of estrangement in a corpus using unique lexical compounds as a measure, it is hoped that a further discussion can take place on what other factors (e.g. distribution, frequency of occurrence) might complement unique LCs in producing an estranging effect, as well as on the practicality of applying linguistic estrangement markers to multimodal and primarily audiovisual media.

References:

Adams, M. (2017). The pragmatics of estrangement in fantasy and science fiction. In M. A. Locher, & A. H. Jucker (eds.), *Pragmatics of Fiction* (pp. 329-63). Berlin: De Gruyter Mouton.

Brezina, V. (2018). *Statistics in Corpus Linguistics.* Cambridge: CUP.

Futurama Episode Scripts. Retrieved from Springfield! Springfield!: https://bit.ly/2RRXM8w (accessed Nov 2019).

Gabrielatos, C. (2018). Keyness Analysis: nature, metrics and techniques. In Taylor, C. & A. Marchi (ed.), *Corpus Approaches to Discourse: A Critical Review*. Oxford: Routledge.

Kilgarriff, A. (2009). Simple maths for keywords. In *Proceedings of the Corpus Linguistics Conference*, Liverpool, Ju

3) Papers at the main conference

Appleton, Stephen

Using corpus techniques to explore lexical change in parliamentary discourse

Democratic parliaments across the world publish transcripts of their proceedings online, which serve both as an official record and as a means of ensuring transparency and accountability. Parliamentary transcripts are a valuable resource for corpus linguistics provided we recognise their limitations: for example, because parliamentary discourse differs from other discourse types, and because the transcripts may not be fully accurate or verbatim (Slembrouck 1992; Mollin 2007).

Working within these limitations, this paper explores how corpus techniques can be used to understand how a parliamentary discourse changes over time. It focuses on lexical change in the foreign policy discourse of the United Kingdom government between 1989 and 2014, a period with major developments in European policy at either end. It uses a bespoke 17-million word corpus assembled from ‘Hansard’ (Parliament.UK, 2019), the Official Report of proceedings in the UK Parliament, comprising the words attributed to ten successive Foreign Secretaries and the junior ministers working with them.

After compiling a complete wordlist for the corpus, the research used frequency analysis to identify which words and phrases display the largest rise or fall in frequency during the period. From this, several themes were selected for detailed investigation. One such theme is ‘freedom’, which is used as a case study in this paper. The word ‘freedom’ rose steadily in frequency through the period, more than quadrupling between the first and last five years. The discourse refers to many kinds of freedom including academic, journalistic, political and religious freedom. Analysis of collocates and concordances reveals which of these grew to form a distinct strand of the UK government’s foreign policy discourse. This in turn provides insights into ministers’ changing conceptualisations of the nature and purpose of foreign policy.

References:

Mollin, S. (2007) ‘The Hansard hazard: gauging the accuracy of British parliamentary transcripts’, Corpora, vol. 2, no. 2, pp. 187–210 [Online]. DOI: 10.3366/cor.2007.2.2.187.

Parliament.UK (2020) Hansard Online [Online]. Available at https://hansard.parliament.uk/ (Accessed 27 November 2020).

Slembrouck, S. (1992) ‘The parliamentary Hansard “verbatim” report: the written construction of spoken discourse’, Language and Literature, vol. 1, no. 2, pp. 101–119 [Online]. DOI: 10.1177/096394709200100202

Balfour, James

FREED TO KILL: a corpus-driven analysis of responsibility in British newspaper articles reporting on schizophrenia

Research in psychology and media studies has shown that the British press over-represent people with schizophrenia in the context of violent crime (e.g. Clement and Foster, 2008; Chopra and Doody, 2007). Cross (2014) has also shown there is typically a moralistic element to these stories, with aggressors contradictorily framed as both 'mad' and 'bad'. Moreover, Clement and Foster (2008) show that violent crimes perpetrated by people with schizophrenia are rarely put into context, often neglecting to mention the reasons why the crimes were perpetrated. However, no research to date has explicitly focussed on the language around responsibility in these stories, and the subtle ways in which the press use language to negotiate blame. This paper redresses this gap and crosses methodological boundaries by presenting a new corpus-based approach for examining the representation of blame and responsibility in news texts reporting on criminal behaviour.

The presentation answers the following research question: 'how do journalists use language to negotiate blame in news stories reporting on schizophrenia and violent crime?' I begin by introducing the data: a 15-million-word corpus of British newspaper articles reporting on schizophrenia. Here I discuss issues relating to corpus compilation and preparation. I then proceed to propose an innovative, interdisciplinary method for analysing the language of blame using corpus tools. This involves identifying the 10 most frequent word forms referring to violent crime in the corpus, and then deriving collocates that orient to responsibility. These words are then categorised and analysed qualitatively with reference to the criteria listed in the Path Model of Blame proposed by Malle et al. (2014).

The analysis reveals that journalists often use language in highly subtle ways to negotiate blame in stories reporting on schizophrenia and crime. This includes (1) the tendency to use words referring to desires and intentions, without necessarily mentioning that these are based on false beliefs, (2) the varied use of speech acts to represent command hallucinations, and implications this has on our understanding of the agent's free will, and (3) the use of ambiguous words and phrases such as allowed and freed to kill, which suggests that medical professionals are complicit in crimes.

I conclude by exploring some of the implications of this research, both on stigma towards people with schizophrenia in society, and on corpus-based methodology.

References:

Chopra, A. K. and Doody, G. A. (2007). Schizophrenia, an Illness and a metaphor: Analysis of the use of the term 'schizophrenia' in the UK national newspapers. Journal of the Royal Society of Medicine, 100(9), 423-426. doi:10.1177/014107680710000919

Clement, S. and Foster, N. (2008). Newspaper reporting on schizophrenia: A content analysis of five national newspapers at two time points. Schizophrenia Research, 98(1), 178-183. doi:10.1016/j.schres.2007.09.028

Cross, S. (2014). Mad and bad media: Populism and pathology in the British tabloids. European Journal of Communication, 29(2), 204-217. doi:10.1177/0267323113516734

Malle, B., Guglielmo, S. & Monroe, A. (2014). A Theory of Blame. Psychological Inquiry, 25(2), 147-186. doi:10.1080/1047840X.2014.877340

Bednarek, Monika

Mediated Australian Aboriginal English: A corpus linguistic study

In the last decade or so, there has been a dramatic turnaround in the amount of Aboriginal and/or Torres Strait Islander television characters in Australia, with a rise in mainstream television drama with Indigenous creative control. Viewers thus encounter a range of Indigenous television characters, who vary in their use of Standard Australian English, varieties of Australian Aboriginal English (AAE), and traditional and new Indigenous languages. However, there is a lack of linguistic knowledge of the language varieties that are transmitted to mainstream audiences in this way. A particular focus of this study is on AAE, which has long been recognised in linguistics 'as a valid, rule-governed dialect of English' (Eades 2013: 2). For viewers, mediated AAE can be an important source of information, especially if they do not regularly interact with Aboriginal and/or Torres Strait Islander people. This would be the case for many Australians, and even more so for international viewers of Australian television series that are exported overseas. This talk focuses on three such series: Redfern Now, Cleverman, and Mystery Road. Using lexical profiling analysis (AntWordProfiler; Anthony 2013) in combination with qualitative concordance analysis, I will identify and compare the use of AAE lexis across the three series. Analysis of frequency and distribution (across series and episodes) will pinpoint words that are particularly significant lexical resources in mediated AAE. Such mediated representation in turn has important social consequences, as it draws on and circulates language attitudes and ideologies, which may influence decision-making and behaviour. The results from this study of Indigenous-authored drama will be compared with previous sociolinguistic research on mainstream narrative mass media (cf. Bednarek 2018: 23-28). This will provide novel insights into the impact of creative control on the deployment of linguistic resources.

References:

Anthony, L. 2013. AntWordProfiler (version 1.4.0) [Computer Software]. Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software.

Bednarek, M. 2018. Language and Television Series. A Linguistic Approach to TV Dialogue. CUP. Eades, D. 2013. Aboriginal Ways of Using English. Aboriginal Studies Press.

Blake, John

Visualizing Rhetorical Moves in a Corpus of Scientific Research

Scientists usually aim to disseminate their work in the highest tier of journal possible. Those able to publish in top-tier journals are most likely to secure employment and tenure at prestigious universities. Research abstracts have been described as a gateway to research articles. The first impression of an article by gatekeepers of scientific journals, editors and reviewers is created on reading the accompanying abstract. Should the abstract fail to show the originality, substance or importance of the research, a negative impression has already been formed in the mind of the reviewer, reducing the possibility of acceptance. Scientific research abstracts, therefore, are a high stakes genre that can make or break careers.

Teachers of scientific writing may often generalize and describe the rhetorical structure of research abstracts in four moves, namely introduction, method, results and discussion (IMRD). In many disciplines this may hold true. To investigate whether IMRD applies to the multidisciplinary field of information science, a corpus of 500 research abstracts from top-tier IEEE journals was compiled. The rhetorical moves in each abstract were manually identified and labelled using a tailor-made schema in the UAM Corpus Tool. In this corpus, analysis of five subdisciplines within information science showed that deviations from the expected IMRD were the norm. In fact, each discipline had its own idiosyncratic style. Rhetorical moves may be cycled through (e.g. MRMR) or inverted (e.g. RM).

An online visualizer was created to help writers in each subdiscipline of information science more easily see how research abstracts are organized. The annotated abstracts were housed in cloud storage categorized into five subdisciplines. Regular expressions are used to match the labels. Any matched expressions are colorized using JavaScript. A web interface was created using React, a JavaScript library. The *Move Visualizer* is online and so can be accessed anytime and anywhere. Users can select their subdiscipline and view abstracts. The colour-coded abstracts help users notice the common patterns of usage in their particular subdiscipline. Exploring the visualizations should help increase awareness of the expected generic conventions of rhetorical moves in their subdiscipline. A comment feature was also added to enable users to provide feedback. Preliminary feedback from three small-scale usability studies was positive with teachers and learners both commenting that they discovered patterns of usage that they were unaware of. This approach to visualization can be adapted to other corpora, enabling corpus linguists to share their annotated texts online.

Bernaisch, Tobias and Julia Degenhardt

Apologies in Indian and Sri Lankan English

Corpus-based research into South Asian Englishes has recently mainly focussed on formal structures (e.g. Lange 2012; Bernaisch 2015). Despite notable exceptions (e.g. Funke 2020), empirical pragmatic research in said varieties is still to be undertaken. Studies of a central pragmatic phenomenon – apologies – have so far concentrated on first-language varieties of English (e.g. Fraser 1981). Against this background, the present study describes the realisation of apologies in two South Asian second-language varieties of English and their historical input variety British English. Relevant research questions include:

1. Are apology frequencies sensitive to structural and contextual factors as well as the speakers' sociobiographic backgrounds?
2. Which factors guide the choice of apology forms and how do these factors influence apology choices?

Apologies were extracted from the spoken parts of the British, Indian and Sri Lankan components of the International Corpus of English based on close readings to identify localised forms and a list of common apology realisations (cf. Deutschmann 2006). After data extraction and cleaning, 645 apologies were annotated for

-VARIETY (i.e. British vs. Indian vs. Sri Lankan English),
-GENDER (i.e. female vs. male),
-AGE (i.e. old vs. young),
-GROUPGENDER (i.e. whether speakers in a group had the same gender),
-INTENSIFIER (i.e. whether the apology was intensified),
-SETTING (i.e. private vs. public),
-TYPE (i.e. speech act vs. speech event),
-TOPIC (of the conversation) and
-TYPE-TOKEN RATIO (of the apologiser).

An improvement on traditional random forests (Breiman 2001; Gries 2019) taking interactions between predictors explicitly into account was employed to model apology realisations. While young British females apologise most often and young Indian males the least, TYPE, TOPIC and GROUPGENDER in interaction with VARIETY guide the choices of apology forms. Sorry as opposed to other apology forms is a) cross-varietally more frequent in speech acts than in structurally more complex speech events, which is particularly evident in Sri Lankan English, b) employed with different frequencies across topics in the varieties concerned and c) most frequent when Indian men apologise to women and Sri Lankans to members of the same sex. In sum, frequencies and forms of apologies appear sensitive to structural, contextual and sociobiograpic factors as well as to the speakers' regional backgrounds.

References:

Bernaisch, T. (2015): The Lexis and Lexicogrammar of Sri Lankan English. Amsterdam: John Benjamins.

Breiman, L. (2001): "Random forests". Machine Learning 45, 5–32.

Deutschmann, M. (2006): "Social variation in the use of apology formulae in the British National Corpus", The Changing Face of Corpus Linguistics, eds. A. Renouf & A. Kehoe. Amsterdam: Rodopi. 205–221.

Fraser, B. (1981): “On apologizing”, Conversational Routine: Explorations in Standardised Communication Situations and Prepatterned Speech, ed. F. Coulmas. The Hague: Mouton. 259–273.

Funke, N. (2020): “Pragmatic nativisation of thanking in South Asian Englishes”. World Englishes (early view).

Gries, S.Th. (2019): “On classification trees and random forests in corpus linguistics: some words of caution and suggestions for improvement”. Corpus Linguistics and Linguistic Theory (early view).

Lange, C. (2012): The Syntax of Spoken Indian English. Amsterdam: John Benjamins.

Bernaisch, Tobias and Nina Funke

Intensifiers and downtoners in South Asian Englishes: Corpus-based perspectives on Indian and Sri Lankan English

While the majority of research on intensifiers (extremely, very) and downtoners (slightly, somewhat) has so far focused on their use in first-language varieties such as British English (e.g. Barnfield & Buchstaller 2010; Fuchs 2017), studies on these functional elements in second-language contexts are rare (see Fuchs & Gut 2016 for a laudable exception). Yet, Hofstede's (1991) multidimensional cultural assessment of Britain, India and Sri Lanka ascribes comparatively higher levels of uncertainty avoidance to speakers from India and Sri Lanka in comparison to British speakers. As higher levels of uncertainty avoidance might be expected to lead to an increase in intensifiers and a decrease in downtoners to make statements more convincing, intensifying and downtoning in British, Indian and Sri Lankan English is studied in relation to these research questions:

1. Are there differences in the frequencies of intensifiers/downtoners regarding factors such as age, gender or regional background of the speakers and can they be reconciled with different degrees of uncertainty avoidance (Hofstede 1991) in the three countries concerned?
2. Which factors influence the frequencies of intensifiers/downtoners and how do they affect intensifier/downtoner use?
3. Are there variety-specific intensifiers/downtoners in the varieties studied and is there a common core of intensifiers/downtoners across the regional varieties under scrutiny?

7925 intensifiers and 878 downtoners were extracted from the spoken parts of the British, Indian and Sri Lankan components of the International Corpus of English. The frequencies of intensifiers and downtoners were considered with regard to their VARIETY, sociobiographic (AGE, GENDER) and contextual factors (FORMALITY, TOPIC, TYPE-TOKEN RATIO). The data for intensifiers and downtoners were modelled via an extension of regular random forests (Gries 2019), which integrates interaction predictors to increase the classification accuracy of the models.

In comparison to British English, lower frequencies of downtoners are observed in Indian and Sri Lankan English and can be reconciled with higher degrees of uncertainty avoidance in the South Asian territories (Hofstede 1991), while the comparatively lower frequencies of intensifiers cannot. Both intensifiers and downtoners occur more frequently with medium and high type-token ratios, in informal settings and with female and younger speakers. Furthermore, relatively large common cores of forms of intensifiers and downtoners are shared across the three varieties.

References:

Barnfield, Katie, and Isabelle Buchstaller. 2010. "Intensifiers on Tyneside: Longitudinal Developments and New Trends". English World Wide 31: 252–287.

Fuchs, Robert, and Ulrike Gut. 2016. "Register Variation in Intensifier Usage across Asian Englishes".
In Heike Pichler, ed. Discourse-Pragmatic Variation and Change in English. Cambridge: Cambridge University Press, 185–210.

Fuchs, Robert. 2017. "Do Women (Still) Use More Intensifiers than Men? Recent Change in the
Sociolinguistics of Intensifiers in British English". International Journal of Corpus Linguistics 22: 345–374.

Gries, Stefan Th. 2019. "On Classification Trees and Random Forests in Corpus Linguistics: Some Words of Caution and Suggestions for Improvement". Corpus Linguistics and Linguistic Theory [available ahead of publication].

Hofstede, Geert. 1991. Cultures and Organizations: Software of the Mind. London: McGraw-Hill.

Burgeois, Samuel

On the evolution of parentheticals in journalistic English

My project argues that parentheticals like (1) and (2) are not only on the rise in certain genres of written English (cf. Leech et al 2009: 246), but that they are also being used at an increasing rate in ways that express the subjective and/or intersubjective stances of the authors who use them (cf. Traugott 2010). This paper will concentrate on written journalism, a genre that has been shown to be going through substantial change from late 20th century to today (e.g. Leech et al. 2009: 240), using the COHA and NOW corpora. Furthermore, this study combines a diachronic investigation of the development of these parentheticals as well as a qualitative look to how they are used in written journalism.

1. (1) In contrast, less than a quarter of the residents of Montana, Hawaii **(well, duh)** and Maine wanted to leave their state. (NOW 2014, emphasis added)
2. (2) With the penultimate volume in her planned seven-book saga about a heroic teen wizard, Rowling restores narrative discipline and ties up some loose threads in preparation for her eagerly awaited **(and, alas, still unscheduled)** grand finale. The horcrux of the matter: Rowling is still at the tippy-toes top of her game. (COHA: Magazines, 2005, emphasis added)

In particular, I will concentrate on parentheticals that express the attitudes and viewpoints of the authors in a way that also seeks alignment with their readership (cf. Jaffre 2009 Du Bois 2007) thus leaving phenomena like nominal appositions to the side. For the purposes of this study, I will look especially at bracketed parentheticals starting with *well* and *and*, though other types will be shown for comparison. Furthermore, I will show that these changes in the use of parentheticals are part of the broader shifts occurring in these genres of written English, especially in terms of the ‘colloquialization’, ‘informalization’ and even the ‘densification’ of modern writing (e.g. Leech et al. 2009; Mair 2006). Furthermore, this study will fill in the gaps that currently exist in the research of parentheticals in writing specifically. For example, though past research has discussed how bracketed parentheticals are particular to and increasingly used in writing (e.g. Leech et al. 2009: 246), detailed empirical research looking into how these parenthetical constructions are used and how they are changing in writing is severely sparse in the current linguistic literature.

References:

Du Bois, John. 2007. The Stance Triangle. In Englebretson, Robert (ed.), *Stance taking in Discourse: Subjectivity, evaluation, interaction*, 139-182. Amsterdam/Philadelphia: John Benjamins.

Jaffe, Alexandra. 2009. Introduction: The sociolinguistics of stance. In Jaffe, Alexandra (ed.) Stance: Sociolinguistic perspectives, 3-28. Oxford: Oxford University Press.

Leech, Geoffrey & Hundt, Marianne & Mair, Christian & Smith, Nicolas. 2009. *Change in Contemporary English: A Grammatical Study*. Cambridge: Cambridge University Press.

Mair, Christian. 2006. *Twentieth-Century English: History, Variation and Standardization*. Cambridge: Cambridge University Press.

Traugott, Elizabeth C. 2010. (Inter)subjectivity and (inter)subjectification: A reassessment. In Davidse, Kristin & Vandelanotte, Lieven & Cuyckens, Hubert (eds.), *Subjectification, Intersubjectification and Grammaticalization*, 29-74. Berlin/Boston: De Gruyter Mouton.

Brunner, Thomas

***We don't play politics with it*. The *play*-X-*with*-Y Construction in Varieties of English**

New varieties of English are a test bed for construction grammar, since language change in them typically occurs at the "interface between lexis and grammar" (Schneider 2007: 83). Along such lines, this study sets out to analyse the *play*-X-*with*-Y construction, as in:

1. (1) *The nationalists are* ***playing games with the people of Scotland***. (GloWbE-GB)
2. (2) [...] *while avoiding being caught in the rain (because* ***it plays havoc with their hair****)* (GloWbE-
AU)
3. (3) [...] *the Republicans are* ***playing chicken with the global economy****, so eff 'em*. (GloWbE-US)
4. (4) *But we should not* ***play politics with the issue*** (GloWbE-NG)

Rather than referring to an actual game, the construction evaluates a mode of interacting with a situation, ranging from positive (*play ball with* 'act fairly', OED) to negative and risky (*play havoc, play chicken with*).

We study all 2,127 instantiations of the *play*-X-*with*-Y construction from the British, the American, the Australian, the Nigerian and the Indian components of the Corpus of Web-based Global English (Davies 2013).

- Using multiple distinctive collexeme analyses, we explore differences with regard to the fillers of the metaphorical X-slot. Against the background of an overall high variability between the varieties, Nigerian English stands out by a marked overuse of *politics* and a pattern of preferences not found in any other variety at hand*.*
- In addition, in Nigerian English, the construction is more than twice as frequent as in all of the other varieties.
- On the basis of LNRE models (Evert & Baroni 2017) we document that the productivity of the X slot in a given variety is strongly proportional to its stage in Schneider's (2007) Dynamic Model (cf. Hoffmann 2014, 2019, 2020; author & collaborator 2020).
- We detect differences in the semantic frames of the complements of *with*, which, for instance, range from highly concrete (cf. (1)–(2)) to abstract (cf. (3)–(5)).
The study demonstrates that even an exceedingly rare construction is both subject to complex variety-specific conditions and aligns itself with an abstract sociolinguistic model.

References

Author & collaborator (2020) Davies, Mark. 2013. *GloWbE corpus*. http://corpus2.byu.edu/glowbe/ (1 July, 2014).

Evert, Stefan & Marco Baroni. 2017. *zipfR: user-friendly LNRE modelling in R*. http://zipfr.r -forge.r-project.org/ (November 27, 2017).

Hoffmann, Thomas. 2014. The cognitive evolution of Englishes: The role of constructions in the Dynamic Model. In Sarah Buschfeld, Thomas Hoffmann, & Magnus Huber (eds.), *The evolution of Englishes*, 160–180. Amsterdam & Philadelphia: John Benjamins.

Hoffmann, Thomas. 2019. *English Comparative Correlatives: Diachronic and Synchronic Variation at the Lexicon-Syntax Interface*. (Studies in English Language.) Cambridge: Cambridge University Press.

Hoffmann, Thomas. 2020. Marginal Argument Structure constructions: The [V the Ntaboo-word out of]-construction in Post-colonial Englishes. Linguistics Vanguard 6(1) .

Hollmann, Willem B. 2013. Constructions in cognitive sociolinguistics. In Thomas Hoffmann & Graeme Trousdale (eds.), *The Oxford Handbook of Construction Grammar, 491–509.* Oxford & New York: Oxford University Press.

Schneider, Edgar W. 2007. *Postcolonial English: Varieties around the world*. Cambridge: Cambridge University Press.

Busse, Beatrix, Ingo Kleiber, Sophie Du Bois and Nina Dumrukcic

Crossing the boundary of time: Retraining modern NLP models for specialized historical corpus data

The application of deep learning (DL) within NLP has yielded promising results for a variety of tasks, and the field has seen a 'neural turn'. While DL approaches have become the standard for contemporary English, historical data has not received the same amount of attention since state-of-the-art models are almost exclusively trained on contemporary language data.

In the spirit of this conference's theme, "crossing boundaries," this paper serves as a case study in how adapting current DL language models to the (historical) corpus domain can improve next-word prediction and additional downstream tasks for working with historical data.

Therefore, the baseline performance of state-of-the-art language models, e.g., BERT (see Devlin et al. 2018) and GPT-2 (see Radford et al. 2019), are compared to models fine-tuned on both our own corpus of 16th-century English grammars as well as external historical data like EEBO TCP (Early English Books Online (EEBO) TCP).

The corpus introduced and utilized in this paper, which is part of the larger HeidelGram project (see e.g. Busse et al. 2020), represents what we label to be British grammars of English from the 16th century. For instance, William Bullokar's Brief Grammar for English, defining amongst other things parts of speech and grammatical cases, constitutes such a grammar.

By applying fine-tuned language models, we are approaching multiple problems identified within the HeidelGram research project, such as mitigating bad scans and OCR as well as diverse classification tasks (e.g., categorizing reference types), from a computational perspective.

In addition, we will propose and demonstrate a relational database approach powering both our corpora as well as our processing and analysis pipelines. While relying on text files and XML annotations (e.g., TEI-based) has been the standard in corpus construction, approaches using relational databases are gaining more attention (e.g., Davies 2005) due to their wide range of advantages (ibid.).

While the focus will be on our specialized grammar corpus, we believe that insights from our experiments, both in terms of fine-tuning existing models as well as using relational databases, will be of interest to everyone working on historical corpora wanting to practically apply current methods from NLP.

References:

Bullokar, William. 1586. Brief Grammar for English. London: Edmund Bollifant.

Busse, Beatrix, Kirsten Gather und Ingo Kleiber. 2020. „A Corpus-Based Analysis of Grammarians' References in 19th-Century British Grammars." In Variation in Time and Space: Observing the World Through Corpora, hrsg. von Anna Cermakova und Markéta Malá. Diskursmuster - Discourse Patterns 20. Berlin: De Gruyter.

Davies, Mark. 2005. "The Advantage of Using Relational Databases for Large Corpora: Speed, Advanced Queries, and Unlimited Annotation." International Journal of Corpus Linguistics 10 (3): 307–34.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee und Kristina Toutanova. 2018. „BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." Unveröffentlichtes Manuskript. https://arxiv.org/pdf/1810.04805.

Early English Books Online (EEBO) TCP. (n.d.): Retrieved November 24, 2020, from https://textcreationpartnership.org/tcp-texts/eebo-tcp-early-english-books-online/

HeidelGram Project: https://heidelgram.de

Radford, A., Jeffrey Wu, R. Child, David Luan, Dario Amodei and Ilya Sutskever. 2019. "Language Models are Unsupervised Multitask Learners."

Bychkovska, Tetyana

Effects of Explicit Instruction on Noun Phrase Production in L2 Undergraduate Writing

Research on syntactic complexity in academic writing revealed that the frequency and complexity of noun phrases increase as writers advance in their academic level or produce higher-quality writing (e.g., Casal & Lee, 2019; Parkinson & Musgrave, 2014; Staples et al., 2016). As a result, scholars increasingly recommend explicit teaching of this grammatical structure in composition classrooms in order to assist students in developing the complexity of their writing more effectively. It is unclear, however, whether such instruction has an effect on writers' use of phrasal features since little research has been done to examine this empirically. Given that writers' syntactic complexity development occurs without explicit instruction and that important elements of existing curricula would likely have to be removed to allow for noun phrase-related activities, it is important to examine the effectiveness of such instruction. To address this gap while also considering effects of prompts on complexity feature production, which were found important in previous research (e.g., Yang et al., 2014), this corpus-based study addresses the following research questions:

1) Does explicit instruction on complex noun phrases in an L2 undergraduate composition course lead to an increased frequency of noun modification types associated with more advanced stages of syntactic complexity development?
2) Does this instruction lead to an increase in the frequency of the total number of complex noun phrases?
3) How does the effect of this instruction differ when the influence of prompt topics is reduced?

Data consist of 120 timed source-based essays from an L2 First-Year Writing course: 60 by the group that received explicit instruction on complex noun phrases and 60 by the group that did not. Within each group, half of the essays that were produced at the beginning of the semester were compared to the other half written at the end. Instruction included explanations of grammatical form and functions, noticing activities, and clause-to-phrase transformations. Complex noun phrases (i.e., noun phrases with at least one element of pre- and/or postmodification) were identified by searching noun modification types from the developmental framework by Biber et al. (2011, pp. 29-30) in the tagged sub-corpora. Frequencies of specific modification types as well as complete noun phrases in each text (rather than the whole corpus) were normed per thousand words, and paired samples t-tests were used for the statistical analysis. Topic effects were determined by conducting a second analysis in which complex noun phrases appearing in prompts were excluded. The results demonstrate that instruction led to more frequent use of complex noun phrases in general and of noun phrase modifiers associated with advanced stages of syntactic complexity development. Also, while prompt topics had an effect on some nominal features, overall trends in development remained unchanged. These findings provide additional evidence for the validity of pedagogical implications suggested in previous research.

Castro Chao, Noelia

Minor complementisers in the temporal domain: The case of English *till* and *until*

Previous research has shown that certain originally adverbial subordinators, such as *as if* in (1), may acquire a complementiser function over time, thereby coming to serve as (near-)equivalents of the declarative complementiser *that*.

(1) *It seemed **as if / that** he was trying to hide his true identity*.

López-Couso & Méndez-Naya (2012, 2015, among others) have extensively discussed the complementiser use of a number of adverbial links which have followed this course of development, for instance *as if*, *as though* and *lest*. The authors show that these so-called 'minor' complementisers typically originate in subordinating links introducing clauses of Comparison (*as if, as though*) and Negative Purpose (*lest*), among others. The change under discussion has been interpreted as a case of secondary grammaticalisation, illustrating a process of "increased grammaticalization of already grammatical items in specific contexts" (Hopper & Traugott 1993: 167; López-Couso & Méndez-Naya 2015: 193–196).

This presentation addresses the same phenomenon in an adverbial domain not explored to date, namely, the domain of Time (Kortmann 1997: 84–85). More specifically, I will consider the history and use of two temporal subordinators, *till* and *until*, which are attested in complementiser function with the verb of Desire *long*, as exemplified in (2).

(2) most marry'd women long **till it be night,** but, for my part, i hate the thoughts of **it** (1681, EEBO BYU)

The study draws on data from a number of sources, including *Early English Books Online* (1470s–1690s; Davies 2017) and the *Corpus of Late Modern English Texts*, version 3.0 (1710–1920; De Smet et al. 2013). Results show that *till/until*-clauses in complement function occur relatively frequently in Early Modern English. However, during Late Modern English, they enter into competition with the *for...to*-infinitive pattern (e.g. 1844, *She longed **for** the old dark door **to close** upon her*), which at the time was beginning to emerge in object position, as documented by De Smet (2013: 90), and are eventually ousted by it. The presentation traces the stages in this process of replacement, and the factors triggering it.

References:

Davies, Mark. 2017. *Early English Books Online*. Part of the SAMUELS project. https://www.english-corpora.org/eebo/.

De Smet, Hendrik. 2013. *Spreding Patterns: Diffusional change in the English system of complementation*. Oxford: Oxford University Press.

De Smet, Hendrik, Hans-Jürgen Diller & Jukka Tyrkkö. 2013. *The Corpus of Late Modern English Texts*, version 3.0. Leuven: KU Leuven.

Hopper, Paul J. & Elizabeth C. Traugott. 2003. *Grammaticalization*. Cambridge: Cambridge University Press.

Kortmann, Bernd. 1997. *Adverbial subordination: A typology and history of adverbial subordinators based on European languages*. Berlin: Mouton de Gruyter.

López-Couso, María José & Belén Méndez-Naya. 2012. On the use of *as if*, *as though* and *like* in Present-day English complementation structures. *Journal of English Linguistics* 40(2), 172–195.

López-Couso, María José & Belén Méndez-Naya. 2015. Secondary grammaticalization in clause combining: From adverbial subordination to complementation in English. *Language Sciences* 47, 188–198. OED = *Oxford English Dictionary Online*. https://www.oed.com.

Collins, Luke

Examining “doctor talk” in a corpus of Emergency Department (ED) interactions

This work presents a corpus-based investigation of 'doctor talk' in the context of Emergency Department (ED) interactions. We evaluate the characteristic features of such talk to determine what 'doctor talk' comprises in an authentic healthcare setting and relate our findings to aspects of communication outlined in instructive texts for doctor-patient communication, namely the Calgary-Cambridge guides (Silverman, Kurtz & Draper, 2013).

The ED Corpus comprises 72 patient journeys as they interact with various healthcare professionals in emergency departments in New South Wales and the Australian Capital Territory. Using the UCREL Semantic Analysis System (Rayson, 2003), we perform keyness analyses to highlight prominent features of 'doctor talk' in comparison with other participants in the interactions. We also compare the language of junior doctors and senior doctors. In addition to highlighting the different roles that junior and senior doctors occupy in the ED, our corpus-based findings indicate how interactional styles differ according to seniority. We identify meaningful differences with respect to aspects of medical talk, information gathering, references to time and space, and discourse structure, which are manifest in the use of lexis related to treatment, questions, deixis and pragmatic markers.

We discuss our findings in comparison to the recommendations for effective clinician-patient interaction (e.g. Silverman, Kurtz & Draper, 2013; Slade et al. 2015), reflecting on the contribution that corpus-based analyses of authentic healthcare interactions can make to assessing and training clinical professionals with respect to effective communication.

References:

Rayson, P. 2003. Matrix: A statistical method and software tool for linguistic analysis through corpus comparison. Ph.D. thesis, Lancaster University.

Silverman, J., Kurtz, S. & Draper, J. 2013. Skills for communicating with patients. Third Edition. Boca Raton, FL, U.S.A.: CRC Press.

Slade, D., Manidis, M., McGregor, J., Scheere, H., Chandler, E., Stein-Parbury, J., Dunston, R., Herke, M. and Matthiessen, C. M. I. M. 2015. Communicating in Hospital Emergency Departments. Heidelberg: Springer.

Correia Saavedra, David, Jennifer Rains and Martin Hilpert

English back-clippings and fore-clippings: What can we learn from corpus data that the dictionaries don't tell us?

English clippings are often categorized into two main types: back-clippings, where the end of the source word is clipped (*bro<brother*), and fore-clippings, where the beginning is clipped (*rona<corona*). While it has been observed that back-clippings are much more common than fore-clippings (Tournier 1985, Lappe 2007, Jamet 2009), an open question is how the choice between back-clipping and fore-clipping can be predicted (Berg 2011, Arndt-Lappe 2018). To shed light on this issue, we have established a database of more than 2000 English clipped words.

Using a stepwise binary logistic regression, we will illustrate how multiple variables influence the choice between these two main types of clippings. We consider well-studied phonological variables, such as the number of syllables, stress, and whether clippings end in a consonant or a vowel (Lappe 2007). In order to determine what can be learned from corpus- based variables, we further consider measures that have received less attention so far, such as token frequency, the number of homographs (e.g. *sub<submarine* vs. *sub<subordinate*), and neighbourhood density (i.e. how many words are "one letter away" from a given clipping, such as *bae<babe* and the word *bane*).

Our results indicate that fore-clippings tend to end in a consonant and have more syllables than back-clippings, while back-clippings originate from longer source words. The corpus- based variables give rise to three observations. First, we observe that fore-clipping favours source words with higher token frequencies. Berg (2011) suggests that the end of a word is less easily recognisable than its beginning, so that fore-clippings should originate from more easily recognisable source words to compensate. Our findings are in line with that suggestion, since high-frequency words are more easily recognized. Second, we find that fore-clippings have fewer homographs than back-clippings. Also this tendency can be explained in terms of recognisability. Third, neighbourhood density did not emerge as a significant variable in our model. We argue that this is the case because neighbourhood density correlates strongly with the number of homographs.

In summary, our analysis suggests that the study of English clippings can benefit from an approach that takes corpus data into account and that tries to model the outcome of the clipping process in terms of information from dictionaries that is combined with corpus-based measures.

References:

Arndt-Lappe, Sabine. 2018. Expanding the lexicon by truncation: Variability, recoverability, and productivity. In Sabine Arndt-Lappe, Angelika Braun, Claudine Moulin & Esme Winter-Froemel (eds.), *Expanding the lexicon: Linguistic innovation, morphological productivity, and ludicity*, 141–170. Berlin: De Gruyter Mouton.

Berg, Thomas. 2011. The clipping of common and proper nouns. *Word Structure* 4(1). 1–19.

Jamet, Denis. 2009. A morpho-phonological approach of clipping in English: Can the study of clipping be formalized? *Lexis – E-Journal in English Lexicology*. HS1. 15–31.

Lappe, Sabine. 2007. *English prosodic morphology*. Dordrecht: Springer.

Tournier, Jean. 1985. *Introduction descriptive à la lexicogénétique de l'anglais contemporain*. Paris-Genève: Champion-Slatkine.

Crosthwaite, Peter

Integrating corpus use into trainee EFL teachers' lesson planning: Building corpus- focused TPACK for the future.

Introduction

The use of corpora for language teaching/learning, via teacher-prepared corpus-assisted materials development or learners' direct use of corpus query software (commonly known as "data-driven learning", DDL) is gaining in popularity in pre-tertiary EFL contexts. However, improving trainee English teachers' technological and pedagogical content knowledge (TPACK) regarding integration of corpus tools/DDL pedagogy into classroom practice has received little attention from a language teacher education perspective. This qualitative study therefore reports on a DDL lesson planning intervention for pre-service secondary school EFL teachers in Indonesia.

Research questions

RQ1) How do trainee language teachers integrate DDL into their lesson planning following DDL training?

RQ2) Do the trainees' LPs demonstrate appropriate TPACK required for successful future implementation?

Approach and Method

Nine pre-service EFL teacher trainees were enrolled in a teacher education program in Jakarta, Indonesia. The DDL training regimen included partial completion of a Short Private Online Course on DDL covering basic corpus techniques required for DDL (e.g. generating corpus queries, reading/manipulating concordances, understanding frequency information) using SKELL (Baisa & Suchomel, 2014) and SketchEngine (Kilgariff et. al, 2014). Trainees then submitted a sample lesson plan which was scrutinized for components where corpus data could enhance the proposed lesson's materials or where learners could engage in direct corpus consultation/DDL. Three, three-hour workshops on DDL were then conducted online via Zoom. Following these, trainees discussed their DDL training and lesson planning via Google Classroom chat, before working alone to create a new lesson plan including at least one direct DDL activity.

Data

Data includes the researcher's initial feedback regarding integration of DDL into trainees' original (non-DDL) lesson plans, trainees' Google Classroom chat logs, and trainees' completed lesson plans involving DDL resources/activities. Harris et. al's (2010) Technology Integration Observation Instrument was used to evaluate trainees' completed lesson plans for TPACK regarding *curriculum goals and technologies; instructional strategies and technologies*; *Technology selection(s)*; and 'Fit'.

Results

Trainees each integrated corpora/DDL into their lesson planning despite none reporting using a corpus prior to training. Submitted lesson plans featured DDL for language-related concerns (e.g. 'grammar focus'), and to support task-based genre-focused pedagogies as required by the Indonesian national curriculum. While submitted plans demonstrated high levels of 'fit' regarding curriculum goals and technology selection, some plans lacked DDL-relevant instructional strategies. However, TPACK scores for submitted lesson plans were generally high following only a short (but intensive) period of DDL training, underscoring the significant potential for integrating DDL into pre-tertiary classroom practice.

References:

Baisa, V. & Suchomel, V. (2014) SkELL: Web interface for English language learning. In Horák, A. & Rychlý, P. (ed.), *Proceedings of Recent Advances in Slavonic Natural Language Processing*. Karlova Studánka, Czech Republic, 5–7 December, 63–70.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, *1*(1), 7-36.

Dayter, Daria and Thomas C. Messerli

Persuasive register in Change My View subreddit

The subreddit R/CHANGEMYVIEW (CMV) is a forum where users come to post an opinion and invite others to change their view in a civilised, well-argued debate. The main persuasion capital for comments on CMV is a special validation system called delta, which can only be obtained if the commenter successfully achieves a 'change of view' of the original poster. Comments which achieve this goal are called DACs (delta awarded comments), whereas all other comments are non-DACs. Due to the delta system, CMV makes for a unique data source where argumentative threads have been pre-annotated by the participants as successful or unsuccessful, and a CMV corpus thus allows the comparative study of persuasive discourse that has persuaded, or has failed to persuade.

For this study exploring persuasive register on CMV, I collected a corpus of all comments posted on CMV between May 2013, when the first content appears, and May 2020, when the data were collected. The data were accessed via a Python script and through the pushshift.io Reddit API. Subsequently, comments tagged as deleted on the subreddit were removed, as were those with a warning that they were in breach of subreddit rules. The current study is based on two smaller subcorpora: the Delta subcorpus, which consists of a 1 million word random subsample of DACs; and a comparable non-Delta subcorpus, which contains a 1 million word subsample of non-DACs.

The conducted study is twofold. First, I used Biber's (1988) register analysis approach to obtain Dimension 4 ('overt persuasiveness') scores for the two subcorpora and to place them on the register continuum relative to the other genres described by Biber. This was done with the help of Nini's (2019) MAT tagger, which replicates Biber's tagging and subsequent factor analysis. The results indicate that both CMV subcorpora are similar to the genre of academic prose, but the DAC subcorpus exhibits higher over persuasiveness score than the non-DAC.

Secondly, I took a closer look at the markers of overt persuasive stance using the corpus-based discourse analysis methodology. I chose two linguistic features that are weighty on Biber's Dimension 4: suasive verbs and prediction modals. For these, I obtained collocation lists from the DAC subcorpus and examined concordances of the highest-ranking collocates. Although this part of the study is still underway and will only be completed in 2021, preliminary results indicate that in many cases, these features do not actually represent an overt persuasive stance. Instead, they appear in the users' meta-discussions about the aim and rules of the subreddit or the meaning of the suasive verbs.

References:

Biber, Doug. 1988. Variation across Speech and Writing. Cambridge: Cambridge University Press. Nini, Andrea. 2019.

"The Multi-Dimensional Analysis Tagger." In Multi-Dimensional Analysis: Research Methods and Current Issues, ed. by Tony Berber Sardinha, and Marcia Veirano Pinto, 67– 94. New York: Bloomsbury Academic.

Del Fante, Dario

Figurative language and pandemics: Spanish Flu and Covid-19 in Us newspapers. A case-study.

The international outbreak of Covid-19 has radically changed our lives and challenged the stability of our contemporary societies, causing a growing distrust of our neighbours. The high intensities of world population mobility have strongly impacted the spread of the virus (Yang et al. 2020). However, this is not the first time that humanity is facing a global pandemic. Almost 100 years ago, large movements of people around the globe in the aftermath of World War I contributed to the spread of an influenza virus that led to one of the most lethal pandemics on record: the 1918 Spanish Flu pandemic.

Even though the divergences between the contemporary and the 1918-19 socio- political contexts, 1918 flu pandemic and coronavirus share basic similarities in the way they're transmitted via respiratory droplets and the surfaces they land on. Metaphors plays a fundamental role in understanding and influencing how we think and talk about health, illness and medicine: each metaphor foreground some aspects of the topic and backgrounds others, and therefore influences how we act, collectively and as individuals (Semino 2008). Since the beginning of the pandemic, metaphor has been widely used to refer to the virus all around the world: both the media and political leaders have often made use of expressions representing the virus as an 'invisible enemy', as 'tsunami on health services', a 'marathon to be endured'. Considering that public opinions are reflected in news and that newspapers are important influencers of people's perspectives on reality, by analysing the metaphorical representation of two pandemics in newspapers in two different periods, we might define, to some extent, how two pandemics are experienced and conceptualised over time. With an understanding of how Covid-19 and Spanish Flu are metaphorically represented in newspaper discourse, it would be easier to shed light on linguistic process through which metaphors work and to understand whether the metaphors used to describe Covid-19 are unique or rather a common feature of communication during pandemics. In order to address this issue, I intend to embark on a case study with a diachronic perspective: drawing on pandemic metaphor research and building on conceptual metaphor theory (Kövecses 2020), this article uses Critical Metaphor Analysis methods (Charteris-Black 2004) to examine and move toward understanding the conceptual metaphor surrounding Covid-19 and Spanish flu. The dataset consists of two broadsheet newspaper corpora: *SpanishFlu Corpus*, composed of articles from *New York Herald Tribune* and *The Evening World*, *CoronaV Corpus*, composed of article from The *New York Times*. These respectively cover the periods 1918-20 and 2019-2020. USA has been chosen because it shows the highest number of Covid-19 infected and because the first confirmed cases of Spanish Flu originated in the United States (Crosby 2003). Preliminary results show that war metaphors are consistently present in both time periods. Disaster metaphors, epicentre of pandemics, are also present in both time period. Fire metaphors are particularly used for Covid-19 but not for Spanish Flu.

References:

Charteris-Black, J. (2004). *Corpus approaches to critical metaphor analysis.* Basingstoke, UK and New York, NY: Palgrave-MacMillan.
Crosby AW (2003). *America's Forgotten Pandemic: The Influenza of 1918* (2nd ed.) Cambridge: Cambridge University Press.

Kövecses, Z. (2020). *Extended Conceptual Metaphor Theory*. Cambridge: Cambridge University Press.Semino, E. (2008). *Metaphor in Discourse.* Cambridge: Cambridge University Press. Webel, M. & Freeman, M. C. (2020) "Compare the flu pandemic of 1918 and COVID-19 with caution – the past is not a prediction." 4th June 2020. <https://theconversation.com/compare-the-flu-pandemic-of-1918-and-covid-19-with-caution-the-past-is-not-a-prediction-138895. Accessed 15/10/2020.

Yang, H., Chen, D., Jiang, Q., & Yuan, Z. (2020). High intensities of population movement were associated with high incidence of COVID-19 during the pandemic. *Epidemiology & Infection*, *148*.

Denison, David, Nuria Yáñez-Bouza, Tino Oudesluijs, Cassandra Ulph and Christine Wallis

Transcribing and Editing *The Mary Hamilton Papers (c. 1750 – c.1820)*

The application of social network theory in English historical sociolinguistics has been a fruitful endeavour (cf. Milroy 1987, Bergs 2005, Sairio 2009). However, despite an increase in social network studies over recent decades, often focussing on rich but relatively small datasets from the eighteenth century (e.g. Fitzmaurice 2000, Tieken-Boon van Ostade 2008, Sairio 2009), much work remains to be done, and many collections are waiting to be transcribed (e.g. *The Elizabeth Montagu Letters*, *The Collected Letters of Hannah More*). In this paper we present our ongoing project on the correspondence and diaries of Mary Hamilton (1756-1816), which will facilitate nuanced social network analyses of language variation and change by means of a highly detailed level of coding of c. 1 million words.

Mary Hamilton was recruited as a royal sub-governess in 1777 and, after leaving the Court in late 1782, she became a member of the Bluestocking circle, counting Hannah More and Frances Burney among her literary friends. The edition of *The Mary Hamilton Papers* will thus provide new insights into the day-to-day life of the royal household in Georgian England and of the artistic and social elites during a period of rapid change in British political, economic, and cultural life. We are currently producing a unified and freely accessible online edition of letters sent to her, drafts/copies by Hamilton herself, a substantial body of her diaries, and a selection of her husband's journals. The output will feature high-quality digital images that will be displayed alongside edited, searchable transcriptions with linguistic and XML/TEI mark-up. In addition, we will offer a comprehensive person database of the correspondents as well as of other people mentioned.

In this paper we will report on the contents of the linguistic corpus and describe our editorial practices and the 'personography' architecture. We will also outline the project research questions addressing three main strands: 1) reading practices, 2) norms and usage, and 3) language structure. The rich materials in the Hamilton archive will allow us to determine whether a social network operates differently in each strand and the extent to which we can generalise across domains.

References:

Bergs, Alexander. 2005. *Social networks and historical sociolinguistics: Studies in morphosyntactic variation in the Paston letters (1421-1503)*. Berlin: Mouton de Gruyter.

Fitzmaurice, Susan. 2000. Coalitions and the investigation of social influence in linguistic history. *European Journal of English Studies* 4(3): 265-76.

Milroy, Lesley. 1987. *Language and social networks*. New York: Blackwell.

Sairio, Anni. 2009. *Language and letters of the Bluestocking network: Sociolinguistic issues in eighteenth-century epistolary English*. Helsinki: Société Néophilologique.

The Collected Letters of Hannah More. http://hannahmoreletters.co.uk/Letters/

The Elizabeth Montagu Letters. http://www.elizabethmontagunetwork.co.uk/

Tieken-Boon van Ostade. 2008. Letters as a source for reconstructing social networks: The case of Robert Lowth. In Marina Dossena & Ingrid Tieken-Boon van Ostade (eds.), *Studies in late modern English correspondence: Methodology and data*, 51-76. Bern: Peter Lang.

Ebeling, Jarle

120 years of reporting clauses: stability or change?

This paper investigates reporting clauses in British fiction over the last 120 years. The focus is on the placement, content and use of the reporting clause. A reporting clause is the narrative part of an orthographic sentence, which also includes quoted or direct speech, i.e. a reported clause. In 1), the reporting clause is *A woman's voice said*, while the rest is the reported clause:

1) A woman's voice said, "Hallo, John, fancy seeing you here."

A reporting clause has two obligatory elements: the speaker/voice and a reporting verb. The reporting verb par excellence is *say*, which, over the past 120 years, has gained ground at the expense of other, semantically richer verbs, e.g. *cry* or *murmur*.

The aim is to establish whether this change in the selection of reporting verb has led to a change in how often the verb is accompanied by i) glossing phrases, in the form adverbs or prepositional phrases, or ii), is "accompanied by a description of following or simultaneous action" (Čermáková and Mahlberg 2018: 231).

1. i) "Yes," *said Bianca, hurriedly*.
2. ii) "I knew," *said Agatha, coming up to the fire*.

If such a change can be detected, may this be seen as a move towards a more elaborate style of reporting speech, where the narrative plays a more prominent part in telling us e.g. how something is said, using some sort of "periphrastic reporting"? Moreover, can it be tied in with the drift towards colloquialization (Mair 2006) or an oral style in written registers as observed by Biber and Finegan (1989)?

The data is culled from the Corpus of British Fiction and consists of more than 1,000 novels published between 1900 and 2019, approx. 85 million words in total.

A random sample of 100 occurrences of direct speech from each of three periods in the last 120 years (1900-1939, 1940-1979, 1980-2019) shows that *say* has increased from 47 to 68 occurrences per 100 reporting clause. This increase in the use of *say* co-occurs with an increase in the use of only having the obligatory elements (speaker + verb) in the reporting clause, but we also find an increase in the use of speech-modifying adverbs and following/simultaneous actions. Other types of speech-modifying elements (adjective and prepositional phrases) and the placement in the sentence of the reporting clause seem to be stable. Overall, very few reporting clauses precede the reported clause in a sentence.

In the talk, I will report on a larger sample of instances to see if this preliminary exploration of the corpus is representative.

References:

Biber, Douglas & Edward Finegan. 1989. Drift and the evolution of English style: a history of three genres. *Language* 65:3. 487–517.

Čermáková, Anna & Michaela Mahlberg. 2018. Translating fictional characters – *Alice* and *the Queen* from the Wonderland in English and Czech. In Anna Čermáková & Michaela Mahlberg (eds), *The Corpus Linguistics Discourse. In Honour of Wolfgang Teubert*, Amsterdam: John Benjamins, 223– 258.

Mair, Christian. 2006. *Twentieth-century English : History, Variation and Standardization*. Cambridge: CUP.

Facchinetti, Roberta, Silvia Cavalieri and Sara Corrizzato

Compiling a spoken corpus of journalistic interviews: from transcription to pragmatic annotation

Spoken corpora have been placed more and more under the lens of linguistic research over the last few years particularly with reference to the issues pertaining to the recording and transcription of discourse events as well as to their mark-up and annotation (Cermak 2009, Kirk and Andersen 2016, Diemer et al. 2016 among others). Though still fewer in number compared to written corpora, spoken corpora are now diversified and wide-ranging both in structure and in purpose (Love & al. 2017) and some of them even rely on novel ways of compilation including the crowdsourcing of language samples (Adolphs et al. 2020).

Taking stock of current research in the field, the present study intends to illustrate a spoken corpus of journalistic interviews currently under development at the University of Verona, Italy. The corpus collects TV interviews where diplomats and international operators are questioned by journalists in English. The corpus is being compiled with the final aim to analyze the typology of questions posed by interviewers on the one hand and the type of answers provided of their interviewees on the other, with special attention to diplomats.

With this aim in mind, a set of metadata have been considered for tagging, including elements concerning both the speakers' identity (i.e., their nationality, profession, nativeness and gender), and the communicative events (i.e., interview date, format, duration and broadcasters). Special attention is dedicated to the annotation of questions and answers to study the interlocutors' stylistic and pragmatic choices and how such choices impact on communication. Specifically, prominence is given to yes-no questions, open questions and alternative questions in relation to the expected answers (Aarts 2018). Directness and indirectness are also taken into consideration to investigate how interviewers elicit information as well as how interviewees provide their answers to fulfil their respective communicative aims.

While illustrating our work-in-progress project, we will discuss the challenges we are facing in terms of (a) XML conversion, (b) metadata definition, and particularly (c) stylistic-pragmatic annotation. We will also illustrate the preliminary results yielded by a pilot corpus we have compiled with interviews focusing on Covid-19.

References:

Adolphs Svenja, Dawn Knight, Catherine Smith and Dominic Price (2020) "Crowdsourcing formulaic phrases: towards a new type of spoken corpus", Corpora 2020 Vol. 15 (2): 141–168.

Aarts, Bas (2018) English Syntax and Argumentation. London: Palgrave.

Cermak Frantisek (2009) "Spoken Corpora Design: Their Constitutive Parameters." International Journal of Corpus Linguistics 14.1 (2009): 113–123.

Diemer Stefan, Marie-Louise Brunner, Selina Schmidt (2016) "Compiling Computer-Mediated Spoken Language Corpora: Key Issues and Recommendations." International journal of corpus linguistics 21.3: 348–371.

Kirk, John M., and Gisle Andersen. "Compilation, transcription, markup and annotation of spoken corpora." International Journal of Corpus Linguistics 21.3 (2016): 291-298.

Love, Robbie, Claire Dembry, Andrew Hardie, Vaclav Brezina, Tony McEnery (2017) "The Spoken BNC2014: Designing and Building a Spoken Corpus of Everyday Conversations." International journal of corpus linguistics 22.3: 319–344.

Franceschi, Valeria

China for the international traveler. A Corpus-based analysis of early travel guidebooks to China

According to Marjorie Morgan (2001), the "traveling age" as we understand it began in the early Victorian years, when leisure traveling within and outside Britain became increasingly common. It is around the same time that we witness a divergence between the basic genres of "travel account" and "travel guidebook", which become "better defined and more easily distinguishable" (Buzard 1993, François 2012: 72-73). In addition to more practical information and a definite paratext, one of the distinctive characteristics of the modern travel guidebook is that the author's subjectivity, ubiquitous in travel accounts, appears here to be condensed in adjectives (Bertho Lavenir 1999: 61).

This study attempts to analyze the presence of authorial subjectivity in early travel guides to China by looking at attitude markers, in order to determine whether and how the authors "signal their attitude towards both their material and their audience" (Hyland & Tse 2004: 156).

The study will be carried out on a corpus of travel guidebooks to China published between 1866 and 1924, compiled using OCR software ABBY FineReader PDF and investigated with corpus analysis software Sketch Engine (Kilgarriff et al. 2014). A mixed approach will be adopted, employing both quantitative – keyness and frequency analysis – and qualitative methods.

The focus of this study is twofold: first, the recurring themes in the guidebooks will be identified through keyword analysis, using the Project Gutenberg English corpus as a reference. Secondly, Hyland and Tse's interpersonal model of metadiscourse (2004) will be borrowed to investigate authorial stance, focusing specifically on attitude markers such as attitude verbs, sentence adverbs, and adjectives (Hyland 2005: 180)

Results are expected to (a) confirm the presence of practical information, such as transportation and accommodation options and prices, and (b) to highlight an overall positive assessment of the locations described in the guides as well as of the Western-owned hotels and travel services offered. Due to a lingering imperial mindset, evaluation of the local population and of the local guides is expected the bear traces of prejudice.

References

Buzard, J. 1993. The Beaten Track: European Tourism, Literature, and the Ways to Culture, 1800-1918. Oxford: Clarendon Press.

Bertho Lavenir, C. 1999. La Roue et le stylo. Comment nous sommes devenus touristiques. Paris : Odile Jacob.

François, P. 2012. "If It's 1815, This Must Be Belgium: The Origins of the Modern Travel Guide Source." Book History 15: 71-92.

Kilgarriff, A., Baisa V., Bušta J., Jakubíček M. Kovář V., Michelfeit J., Rychlý, P. Suchomel, V. 2014. "The Sketch Engine: Ten Years on." Lexicography 1: 7-36.

Hyland, K., Tse, P. 2004. "Metadiscourse in Academic Writing: A reappraisal." Applied Linguistics 25(2): 156-177.

Hyland, K. 2005. "Stance and Engagement: a Model of Interaction in Academic Discourse." Discourse Studies 7(2), 173-192.

Morgan, M. 2001. National Identities and Travel in Victorian Britain. New York: Palgrave.

Garretson, Gregory and Rachele De Felice

Detecting and analysing problem-oriented language in an email corpus

Identifying and solving problems is a central function of language. What we term Problem-Oriented Language (POL) has been studied primarily in workplaces from linguistic and organizational-management perspectives (where it is frequently termed Problem-Solving Talk, e.g. Kim and Angouri 2019, Mangrum et al. 2001), but the focus is generally on spoken interactions like meetings, where it is identified manually. Here we present a corpus-based approach to detecting such discourse in workplace emails and show how this can expand current theories of POL. We answer these questions: 1) Can we automatically identify discussion of 'problems' in a workplace email corpus? 2) What does analysis of these instances contribute to our understanding of POL?

We used the Clinton Email Corpus (De Felice and Garretson 2018), an invaluable resource for the task for several reasons. It combines a large number of real-world workplace emails (over 33,000) with detailed information about the interactants, as most of their identities are known. Our analysis thus includes information about workplace hierarchy and interpersonal relations. Furthermore, the nature of the corpus means there is a range of problems to study, from scheduling miscommunications to serious political crises and natural disasters.

For question 1), we developed a method using collocational analysis and email metadata. We created a set of search terms from 'problem talk' seed words (e.g. *solve, issue, unfortunately*) based on the literature and augmented with synonyms, and expanded this iteratively by identifying these words' most frequent collocates. Then, a custom corpus search output a frequency count for each problem-related lexeme on a per-text basis. We then plotted the density of problem-words on a timeline based on the emails' timestamps, to identify peaks of activity. The level of analysis can be varied: small daily spikes identify minor problems, while surges over several days identify major geopolitical events.

For question 2), we identified several instances of problems and resolutions and analysed these qualitatively. For each, we examined the language used to describe seven dimensions of the problem (e.g. Cause, Affected, Solution). Interestingly, we found a key linguistic dimension of POL to be discussion of both material and emotional effects of the problem on those affected and those tasked with solving it. In this corpus, we noted a greater focus on identifying solutions than on assigning blame or dispensing criticism. The case studies inform the continued development of both our model of POL and our method of detection.

Our presentation will describe the method developed, present our model of problem-solving talk, and work through some case studies from the corpus.

References:

De Felice, R., and Garretson, G. 2018. Politeness at Work in the Clinton Email Corpus: A First Look at the Effects of Status and Gender. *Corpus Pragmatics (2018)2*: 221-242.

Kim, K., and Angouri, J. 2019. 'We don't need to abide by that!': Negotiating professional roles in problem-solving talk at work. *Discourse & Communication, 13(2)*: 172–191.

Mangrum, F. G., Fairley, M. S., & Wieder, D. L. 2001. Informal Problem Solving in the Technology-Mediated Work Place. *The Journal of Business Communication, 38(3)*, 315–336.

Gilquin, Gaëtanelle and Lea Meriläinen

Constrained communication in EFL and ESL: The case of embedded inversion

Second language acquisition and contact linguistics have mostly been approached as different fields of research. Yet, the past decade has shown that the boundaries between these fields can be crossed thanks to the corpus-based comparison of English as a foreign language (EFL, aka Learner Englishes) and English as a second language (ESL, in the sense of New Englishes). The similarities found between the two types of varieties (see, e.g., Deshors et al. 2016) may be explained by the fact that they are both instances of so-called constrained communication, i.e., language production characterized by certain socio-cognitive constraints (Kruger & Van Rooy 2016). More precisely, EFL and ESL share the constraint of bilingual language activation, in that all EFL and ESL speakers use English as an additional, non-native language.

The effect of this shared constraint is here investigated through embedded inversion (EI), i.e., the inversion of SV word order in indirect questions, as illustrated by "One should ask then at what expense have these benefits been reached" (ICE-Sing). EI represents a non-standard feature which has been found to be attested in many geographically distant ESL varieties (cf. Sand & Kolbe 2010). We ask whether EI can also be found in EFL and, beyond attestation, whether common linguistic realizations appear in EFL and ESL, for example phraseological patterns that tend that favour EI.

Our study relies on several (sub)corpora of EFL and ESL (ICLE, LINDSEI, ICE) which, besides the common bilingual constraint, differ along constraint dimensions like modality (speech/writing) and task expertise (novice/expert writing) (cf. Kotze 2020). Cases of EI are retrieved by manually examining the concordances of 114 verbs taken from Biber et al. (1999). The data is further annotated for linguistic variables that have been found to influence the occurrence of EI, such as chunks with the verb BE and subject length. The significance of the different variables is analysed with the R table.plotter function (Gries 2009). The analysis shows that EI, though relatively rare, is found in both EFL and ESL, and in partly similar contexts (e.g. with the chunk "what's" or in the second part of a coordinated embedded clause). Some constraints appear to play a role in the use of EI, such as the degree of expertise, with EI being more frequent among novice than expert writers.

References

Biber, D. et al. 1999. *Longman Grammar of Spoken and Written English*. Longman. Deshors, S. et al. 2016. Linguistic Innovations: Rethinking Linguistic Creativity in Non-native Englishes. Special issue of *IJLCR* 2(2).

Gries, S. 2009. Table.plotter 0.97. A function for R.

Kotze, H. 2020. Converging *what* and *how* to find out *why*. An outlook on empirical translation studies. In *New Empirical Perspectives on Translation and Interpreting*, L. Vandevoorde et al. (eds), 333-371. Routledge.

Kruger, H. & Van Rooy, B. 2016. Constrained language: A multidimensional analysis of translated English and a non-native indigenised variety of English. *EWW* 37(1): 26-57.

Sand, A. & Kolbe, D. 2010. Embedded inversion worldwide. *Linguaculture* 2(1): 25-42.

Hannaford, Ewan

Interpreting Illness: A corpus linguistic investigation of represenations of mental and physical illness in the UK and US press (1995-2017)

Medical conceptualisations of illness have developed substantially over the past few decades, particularly in desegregating categories of physical and mental illness (Satcher & Rachel, 2017; Stein et al., 2019). However, public attitudes towards mental and physical illness remain distinct, with this particularly evident in the continued stigmatisation of mental health problems. Models of public attitudes and illness representations suggest media portrayals play a pivotal role in the continued alienation of mental health conditions (Young, Norman & Humphreys, 2008; Ross et al., 2019). As the mechanisms by which media discourse constructs public conceptions of illness have previously been under-examined and direct comparisons between coverage of mental and physical illnesses have been scarce, my research project analyses how media illness discourses contribute to public knowledge of mental and physical illness by addressing three key questions:

1. How does media coverage construct key dimensions of illness representations?
2. Do media representations accurately reflect modern medical conceptualisations of health conditions, or has coverage perpetuated a divide between mental and physical illnesses?
3. Are stigmas, stereotypes, or other negative elements attached to specific illnesses, or categories of illness, in media coverage?

This paper focuses on distinctions in coverage of mental and physical health conditions and the persistence of divisions between these categories in the press. A combined analysis using keywords, collocates, and semantic tagging was conducted on two corpora of UK and US newspaper coverage (1995-2017), with key-keywords calculated for press discourse on different illnesses in three distinct periods and then grouped semantically to compare salient themes. Collocational and concordance analysis then further informed investigations of particular key-keywords in context. Through this multifaceted approach, press representations of nine diverse illnesses were compared with their medical conceptualisations, drawing on relevant, contemporary scientific literature surrounding each condition.

My results show some progress in press representations of mental/physical health in line with medical advancements, such as growing cross-categorical overlap in other conditions mentioned in individual illness discourses. However, misconceptions, stereotypes, and inaccuracies are also perpetuated in press coverage, such as a continued association between mental illness and violence/criminality, as well as specific features of illnesses being underrepresented in coverage, such as the psychological impacts of physical conditions. Considering the impact of these discursive features on public interpretations of illness, this paper also suggests pathways by which press representations may better reflect medical understanding of mental and physical illness and reduce stigma attached to specific illnesses.

References:

Ross, A.M., Morgan, A.J., Jorm, A.F., & Reavley, N.J. 2019. A systematic review of the impact of media reports of severe mental illness on stigma and discrimination, and interventions that aim to mitigate any adverse impact. Social Psychiatry and Psychiatric Epidemiology 54(1). 11–31.

Satcher, D. & Rachel, S.A. 2017. Promoting Mental Health Equity: The Role of Integrated Care. Journal of Clinical Psychology in Medical Settings 24(3–4). 182–186.

Stein, D.J., Benjet, C., Gureje, O., Lund, C., Scott, K.M., Poznyak, V., & Van Ommeren, M. 2019. Integrating mental health with other non-communicable diseases. BMJ (Online) 364. 13–16.

Young, M.E., Norman, G.R., & Humphreys, K.R. 2008. Medicine in the popular press: The influence of the media on perceptions of disease. PLoS ONE 3(10). 1–7.

Hannader, Helena

Working towards a gold standard in writing revision analysis

The primary focus of learner corpus research has been on the text as a product. Writing revision analysis allows access to the writing process, which in turn reveals additional information about interlanguage. So far, LCR on writing revisions has focused on advanced learners and very specific topics such as timing in keystroke logging or peer reviewing without having a solid baseline to which these findings can be compared (cf. Lindgren 2005, Razak & Saeed 2015). When studies on intermediate learners were conducted, they were rather small-scale (Kreyer 2019). This limited statistical analysis, so that more broad, potentially generalisable findings are still waiting to be uncovered.

The present paper will attempt to elevate writing revision analysis of intermediate L2 speakers by examining the /Marburg Corpus of Intermediate Learner English/ (MILE, slashes indicate italics). The longitudinal data consists of texts written as part of their exams by 91 secondary school students from grades nine to twelve and cumulates in 15,302 revisions. The exams were handwritten, so that the revised sections could be marked up and included in the digitised corpus. This allows for a process-based approach that adds to the research findings of the usually product-based investigations in corpus research. However, delineating the specific change that is present in a revision can be difficult, which we want to explore. This paper will work towards closing this gap by presenting a taxonomy and rules to guide writing revision analysis generally as well as annotation of the MILE specifically. The observations that will be discussed range from uncomplicated ones to more problematic cases, such as:

(1a) Link left [his family] Bradford (0004-1-09-00109, square brackets mark deleted sections)

(1b) The [wour] world would not understand their goals (0082-2-12-00073)

(1c) […] so it is easier for Iago to [get] make Othello jealous than Desdemo. (0061-2-11-00052)

(1d) Im very happy, that you choosed Hong Kong for your geography [present] presentation (0001-1-09-00002)

The addition in (1a) is a straightforward observation concerning content, while (1b) exhibits a spelling revision. (1c) could be a content revision (target hypothesis could have been /get Othello angry/) or a change for the sake of a collocation. This opens up at least three possibilities: uncertainty between /get jealous/ and /make jealous/, uncertainty concerning an alternative collocation such as /get angry/ and /make angry/ or a change of content. Lastly, the revision in (1d) could be a morphological or a morphosyntactic alteration, depending on whether the student was unsure about the correct noun form or about the word class that was required. This paper will discuss related principles, pitfalls and problems and propose how to address them. In this way, we want to contribute to a gold standard in revision analysis.

References:

Kreyer, R. (2019). "Dear [Man men and women] madam, dear xxx sir"— What we can learn from revisions in authentic learner texts. Corpus Linguistics, Context and Culture. De Gruyter. 63–386.

Lindgren, E. (2005). Writing and Revising: Didactical and Methodological Implications of Keystroke Logging. Umeå Universitet.

Razak, N. A., & Saeed, M. A. (2015). EFL Arab Learners' Peer Revision of Writing in a Facebook Group: Contributions to Written Texts and Sense of Online Community. English Language Teaching, 8(12), 11–26.

Hober, Nicole

Hybrids compounds in World Englishes

This project is contextualised within recent research on lexical creativity and word-formation at the intersection of World Englishes and Second Language Acquisition (Biermeier, 2014; Callies, 2016; Deshors, Götz, & Laporte, 2018) and further examines lexical innovations with a special focus on hybrid compounds. Hybrid compounds are formed by combining an English element and an element of indigenous origin. Prominent examples are frequently cited combinations with Hindi *wallah* 'person (connected with a particular thing or activity)' in Indian English, e.g. *rickshaw wallah* or *plastic wallah*. These hybrid formations are also found in Pakistan English as shown in Example (1).

(1) *The **fruit wallah** had seasonal fruit only.* (GloWbE; Pakistan: General)

I adopt a usage-based, cognitively-motivated, and process-oriented approach to investigating hybrid compounds in 14 ESLs. The general premise motivating this project is the observation that concepts for which no terms in 'native' English varieties exist, but which are relevant to speakers of any ESL variety, "adapt to the sociocultural reality in the country" (Dako, 2001: 26).

The data are taken from the *International Corpus of English* (ICE), tagged for indigenous items, and the *Corpus of Global Web-Based English* (GloWbE), offering a much larger more recent database. 20 indigenous elements found in hybrid formations in ESLs were re-investigated to validate and complement previous findings (cf. Biermeier, 2014) and to discuss the hybrid formations' development. Both type and token frequency lists were compiled. The data was manually annotated for conceptual domains (e.g. FOOD, PEOPLE), morphological combination patterns, and internal semantic structure (compositional vs. non-compositional). Three guiding questions were addressed:

1. What different kinds of hybrid compounds, from a morphological and semantic point of view, can be found in ESLs? Can a taxonomy of the possible combination patterns be generated?
2. Which conceptual domains are 'affected'?
3. What are the underlying processes and motivations for the creation of hybrid compounds?

Over 1,100 hybrid compound types were identified. Compositional two-element hybrid formations where the English item functions as the head are the preferred combination pattern. All 'affected' conceptual domains are essential for everyday life.

Based on the empirical findings and a fine-grained qualitative analysis, I propose that three operating forces or motivating factors, i.e. salience, frequency, and economy, determine why certain words serve as input for hybrids while others do not.

References:

Biermeier, Th. 2014. Compounding and suffixation in World Englishes. In S. Buschfeld, Th. Hoffmann, M. Huber, & A. Kautzsch (Eds.), *The Evolution of Englishes: The Dynamic Model and Beyond*, 312–330. Amsterdam: John Benjamins.

Callies, M. 2016. Towards a process-oriented approach to comparing EFL and ESL varieties: A corpus-study of lexical innovations, *International Journal of Learner Corpus Research 2*(2), 229–250.

Dako, K. 2001. Ghanaianisms. Towards a semantic and a formal classification, *English World-Wide 22*(1), 23–53.

Deshors, S. C., Götz, S., & Laporte, S. (Eds.). 2018. *Rethinking the linguistic creativity of non-native English.* Amsterdam: John Benjamins.

Iyeiri, Yoko

Exploring Benjamin Franklin's English: A Case of Intra-Writer Variation

In response to the growing interest in intra-writer variation in recent historical sociolinguistics, the present study looks into Benjamin Franklin's letters and autobiography, focusing on the variability of English in them. Although studies on particular authors' English are numerous, they tend either to pay little attention to the variability of language or to discuss it mainly from stylistic perspectives. The present study intends to clarify whether it is possible to contextualize Benjamin Franklin's English within the history of the English language. In other words, the aim is to observe different stages of language change within a single person's variation. Hence, the method employed is a variationist one, which is well-established and almost traditional in corpus linguistics. Traditional parameters like dates as in first-generation corpora or social ranks as in second-generation corpora are, however, no longer usable for the purpose of the present study, which focuses on one person's English. Furthermore, intra-writer variation tends to be subtle, since the material to be explored is attributable to a single author in the end. This difficulty is to be surmounted by shedding light on numerous different aspects of language. This study focuses on some morphological, syntactic, and discoursal features of Benjamin Franklin's English.

I have used for this purpose approximately 600,000 words of Benjamin Franklin's texts available as open resources, including his autobiography and selected letters. Concerning the language of particular texts generally, morphological features are easier to identify than syntactic ones. This is due to the fact that morphology tends to be more coherent and stable in limited datasets than syntax (cf. Hogg 2006, p. 356), while the latter can fluctuate at intra-writer or even intra-text levels. Contrary to this general assumption, however, Benjamin Franklin shows variation to some noticeable extent in morphology (e.g. forgot vs. forgotten and writ vs. written), and more significantly, the variation seems to be conditioned by sociolinguistic factors. I will analyze the difference between letters and the autobiography, and within letters between those addressed to Deborah, Benjamin Franklin's wife and those addressed to other people. Together with syntactic and discoursal features, which are more likely to clarify intra-writer variation even at the synchronic level (e.g. the be- and have-perfect of some intransitive verbs, and the use of private verbs), this paper shows that morphological features also provide hints as to the process of ongoing language change.

Reference:

Hogg, Richard. (2006) English in Britain. In Richard Hogg and David Denison (Eds.), A History of the English Language (pp. 352-383). Cambridge: Cambridge University Press.

Kaunisto, Mark and Paula Rautionaho

Was/were variation with plural subjects in recent British English – towards standard uses?

The alternation of *was* and *were* is a linguistic phenomenon that has drawn much attention in sociolinguistic and dialectological studies in recent years. Studies on the variation have observed uses in varying degrees of focus, ranging from surveys examining the variation in specific cities (e.g. Tagliamonte 1998) to those covering nationwide patterns of use (e.g. Wolfram & Schilling-Estes 2003; Hay & Schreier 2004). Different types of regional uses have been identified, such as *was*-levelling, where *was* is used in grammatical contexts where *were* would be normally used (e.g. with plural subjects), as well as *were*-levelling, where the opposite holds. In some varieties, the spread of one form has been seen to occur in specific grammatical contexts, with e.g. *were* being preferred in contexts involving negation (Tagliamonte 1998), a phenomenon which in some dialects is particularly prominent in tag questions (Anderwald 2002).

The paper proposed here explores recent trends in the *was*/*were* variation in recent British English, focussing on the uses of *was*/*were* with plural pronoun subjects in the spoken demographic part of the BNC and the Spoken BNC2014 corpus. Extracting all instances of *was* and a random sample of 500 instances of *were* with plural pronoun subjects in both corpora, we annotated the data for intra-linguistic (e.g. verb form, negation, pronoun) and sociolinguistic (age, gender, region, and social class) variables. While a striking decline in the normalized frequencies of *was* with plural pronoun subjects is undisputable (from 123.5pmw in 1994 to 30.9pmw in 2014), we dig deeper into intra- and extra-linguistic parameters to reveal the changing patterns at hand with generalized linear mixed model tree analysis (GLMM tree; Fokkema et al. 2018). The results indicate that the sociolinguistic parameters override intra-linguistic ones; the major divide is found between upper and lower social classes, the north and the south, and the younger age groups as opposed to the older ones, while pronouns are the only intra-linguistic parameter chosen in the final model. These observed contrasts together with the overall decrease in the use of *was* with plural pronoun subjects give rise to further considerations on the phenomenon from a sociolinguistic point of view, while also addressing questions relating to the nature of the data.

References:

Anderwald, L. 2002. *Negation in non-standard British English: gaps, regularizations and asymmetrics*. London: Routledge.

Fokkema, M., Smits, N., Zeileis, A., Hothorn, T., and Kelderman, H. 2018. Detecting treatment subgroup interactions in clustered data with generalized linear mixed-effects model trees. *Behavior Research Methods*, 50: 2016–2034.

Hay, J. and Schreier, D. 2004. Reversing the trajectory of language change: Subject-verb agreement with be in New Zealand English. *Language Variation and Change*, 16: 209–236.

Tagliamonte, S. 1998. *Was/were* variation across the generations: View from the city of York. *Language Variation and Change*, 10(2): 153–191.

Wolfram, W., and Schilling-Estes, N. 2003. Parallel development and alternative restructuring: The case of weren't regularization. In D. Britain & J. Cheshire (eds.), *Social Dialectology: In honour of Peter Trudgill*. Amsterdam: John Benjamins. 131–154.

Kehoe, Andrew, Matt Gee and Antoinette Renouf

A collocational approach for explaining changes in word frequency over time

Time series analysis has been of growing interest in the study of language change (e.g. Petersen et al. 2011, Grieve et al. 2017). In addition, change in the collocational profiling of a word has seen increased attention (e.g. Sagi et al. 2011, Brezina 2018). Bringing together these two areas, this paper conducts a diachronic study of a corpus covering 30 years of mainstream UK news text.

In our previous studies we presented approaches to finding instances of word frequency change in a data-driven manner, leading to the discovery of word usage fluctuation and frequency change points. To discover significant changes in word frequency over time three tests were employed: Cox's sequential test (Cox 1952), to find trends; shifts in mean over time, to find sudden frequency jumps; and coefficient of variation, to find seasonal patterns. Thresholds were established for the tests where crossing them indicated significant variation.

We now turn our attention to the next phase of analysis and discuss methods employed to account for changes in word frequency. Some fluctuations are easy to explain. For example, the cyclical patterns observed for *Christmas* and *Olympics* are determined by real-world events, and the upward and downward trends for *Instagram* and *iPad* relate to technological changes.

However, the reasons for other changes can be harder to determine, especially where many thousands of examples need to be inspected. To assist in explaining such changes, we adopt a collocational approach. We define collocation as the statistically significant co-occurrence of a pair of words within a fixed span (1 or 4 words). To study this phenomenon over time, we calculate the collocational profile of each word on a month-by-month basis over the 30 year period. We then present the profile using a visualisation known as an horizon graph (Saito et al. 2005). Using this method we observe semantic changes, for example, *bullying* shows increased association with *online* and *cyber* from June 2005 onwards. Meanwhile, the horizon graph for *gender* shows how changes in societal norms are reflected linguistically in new collocates such as *binary*, *non-binary*, *trans* and *fluid*.

We conclude with a discussion of potential refinements. We observe that the frequency of semantically- related collocates can fluctuate in unison, which opens up the possibility of measuring the correlation between time series.

References:

Brezina, Vaclav. 2018. Change over Time. In *Statistics in Corpus Linguistics*, 219–256, Cambridge: Cambridge University Press.

Cox, David, 1952. Sequential tests for composite hypotheses. *Mathematical Proceedings of the Cambridge Philosophical Society*, 48(2) 290-299.

Grieve, Jack, Andrea Nini & Diansheng Guo. 2017. Analyzing Lexical Emergence in Modern American English Online. *English Language and Linguistics*, 21(1) 99–127.

Petersen, Alexander, Joel Tenenbaum, Shlomo Havlin & Eugene Stanley. 2012. Statistical Laws Governing Fluctuations in Word Use from Word Birth to Word Death. *Scientific Reports*, 2(1).

Sagi, Eyal, Stefan Kaufmann, and Brady Clark. 2011. Tracing semantic change with Latent Semantic Analysis. In Allen, Kathryn and Justyna Robinson (eds.), *Current Methods in Historical Semantics*. De Gruyter, 161-183.

Saito, Takafumi, Hiroko Nakamura Miyamura, Mitsuyoshi Yamamoto, Hiroki Saito, Yuka Hoshiya, and Takumi Kaseda. 2005. Two-tone pseudo coloring: Compact visualization for one-dimensional data. In *IEEE Symposium on Information Visualization INFOVIS 2005*, 173-180, IEEE.

Kemble, Melissa

Female athletes and masculine sports: A corpus analysis of patriarchal discourses in the Australian print media

This research is a corpus-based discourse analysis investigating how athletes are represented in the print news media, with respect to patriarchal discourses. Previous research into sports news coverage has identified discursive practices that highlight and reinforce stereotypes of 'appropriate' femininity in sport (Bruce 2016). Much of this existing research, however, sits outside the discipline of linguistics, with only a small portion of studies taking a corpus- linguistic approach. This study focuses on the historically and stereotypically masculine sports of Australian Rules Football (AFL) and Rugby League (NRL). The recent advent (in 2017 and 2018) of the new elite women's competitions into these well-established male sporting domains presents a unique opportunity for linguists to investigate whether historical gender bias in sports news coverage persists today.

To date, only two studies investigate gender bias in media reporting on women's AFL (Kemble 2020; Sherwood et al. 2019), with none focusing on women's NRL. Drawing on Kemble's (2020) research as a baseline, this study significantly expands the scope and analyses a more comprehensive dataset. Specifically, this research investigates how elite female and male AFL and NRL athletes are represented in the Australian print media, with respect to previously documented patriarchal discourses i.e., objectification, trivialisation, stereotyping and othering. A specialised corpus of Australian print news articles (the 'OzFooty' corpus) has been constructed from the five most widely read newspapers in Australia for the period 1 December 2017 2019. The research combines corpus linguistic analysis, using AntConc (Anthony, 2020) and the UCREL Semantic Annotation System integrated in WMatrix (Rayson, 2009) with qualitative discourse analysis of evaluation. Corpus linguistic techniques, including frequency lists, keywords and semantic tags, have been employed to identify salient patterns pointing to gender biases. Text analysis of evaluation has then been undertaken to investigate the attitudes expressed towards the athletes with respect to patriarchal discourses. Initial results indicate a potentially positive shift away from previously documented patriarchal discourses, with female athletes portrayed as "elite sports players" (Caple, 2013, p. 288). However, the text analysis also reveals evidence of underlying and implicit biases which continue to position the AFL and NRL sporting spaces as ideally and preferably male.

This research contributes to the existing literature on gender bias in sports news reporting by highlighting how elite female athletes newly entering a male-dominated sporting space are represented in the media. Additionally, the mixed-method approach combing corpus linguistics with qualitative discourse analysis of evaluation provides a foundation for future linguistic research into representations of athletes in sports news discourse.

References:

Anthony, L. (2020). AntConc (Version 3.5.9) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/

Bruce, T. (2016). New rules for new times: Sportswomen and media representation in the third wave. *Sex Roles*, 74, 361-376. doi: 10.1007/s11199-015-0497-6

Caple, H. (2013). Competing for coverage: Exploring emerging discourses on female athletes in the Australian print media. *English Text Construction, 6*(2), 271-294. doi: 10.1075/etc.6.2.03cap

Kemble, M. (2020). As good as the men? A corpus analysis of evaluation in new articles about professional female athletes competing in 'masculine' sports. *Critical Approaches to Discourse Analysis Across Disciplines, 12*(1), 87-111.

Rayson, P. (2009) Wmatrix: a web-based corpus processing environment. Computing Department, Lancaster University. Available from: http://ucrel.lancs.ac.uk/wmatrix/

Sherwood, M., Lordanic, M., Bandaragoda, T., Sherry, E., and D. Alakakoon (2019). A new league, new coverage? Comparing tweets and media coverage from the first season of AFLW. *Media International Australia,* 172(1): 114-130. doi: 10.1080/09523360008714133

Klavan, Jane and Sandra-Leele Toom

The use of phrasal verbs by Estonian-speaking EFL learners

The aim of the paper is to investigate, by replicating the study by Gilquin (2015), the use of phrasal verbs by Estonian-speaking foreign learners of English. Three research questions are proposed: (a) how often are phrasal verbs used by Estonian EFL learners in written and spoken language; (b) which phrasal verbs are used most frequently by Estonian EFL learners; (c) how does the use of phrasal verbs by Estonian EFL learners compare to the use of phrasal verbs by native speakers. These questions were answered by conducting a corpus-based analysis. Two corpora were used, the Tartu Corpus of Estonian Learner English (TCELE) and the Estonian subcorpus of the Louvain International Database of Spoken English Interlanguage (LINDSEI-EST). Data from the corpora were entered into a concordance program and a lexical search was conducted with 25 different phrasal verb particles.

The use of phrasal verbs by Estonian EFL learners was divided fairly evenly between written and spoken language, with a slightly higher frequency found in spoken language (116 vs 119 phrasal verbs in written and spoken language respectively). The relative frequency of phrasal verbs used by Estonian EFL learners in spoken language was 56 per 10,000 words and in written language, 47 per 10,000. Native speakers, as described in Gilquin's (2015: 60) study, used phrasal verbs more often, with a relative frequency of approximately 100 per 10,000 words in spoken language and 50 per 10,000 words in written language. Compared to native speakers, Estonian EFL learners underuse phrasal verbs, especially in spoken language; the relative frequencies were more similar in written language. The data analysis showed that of the 109 different phrasal verbs identified, the most frequently used phrasal verbs were go on, take over, and sum up. The most frequently used particles were up and out, which were the same for native speakers (Gilquin 2015: 68).

These and future results can highlight the issues Estonian EFL learners face when using phrasal verbs and aid in developing teaching methods to target their specific needs to improve their understanding and use of phrasal verbs. Phrasal verbs, although seemingly simple, are a notoriously difficult topic for EFL learners to grasp (e.g. Dagut & Laufer 1985, De Cock 2006, Gilquin 2015, Hulstijn & Marchena 1989). Despite the fact that phrasal verbs have been widely researched in the field of EFL, there is not much research available in the context of Estonian EFL learner language.

References:

Dagut, Menachem and Batia Laufer. 1985. Avoidance of phrasal verbs: A case for contrastive analysis. Studies in Second Language Acquisition, 7: 1, 73-79.

De Cock, Sylvie. 2006. Learners and phrasal verbs. MED Magazine, 35: January–February.

Gilquin, Gaëtanelle. 2015. The use of phrasal verbs by French-speaking EFL learners. A constructional and collostructional corpus-based approach. Corpus Linguistics and Linguistic Theory, 11: 1, 51-88.

Hulstijn, Jan and Elaine Marchena. 1989. Avoidance: grammatical or semantic causes? Studies in Second Language Acquisition, 11: 3, 241-255.

Kohn, Birgit

When ‘very crazy’ is not crazy enough: Creative ADJ-intensifying constructions

It is argued that creative expressions deviate from or extend beyond conventional word use (Hanks 2013; Goldberg 2019). Based on this, analysing creative expressions needs to be contrasted with conventional word use in the sense of highly frequent, entrenched patterns. Assuming that entrenchment is closely linked to corpus frequency and to productivity (Stefanowitsch & Flach 2017; Perek 2018), statistical measures such as collostructional analysis (Stefanowitsch & Gries 2003) and different measures of productivity offer tools to determine a construction’s level of productivity and the entrenchment of individual constructs. Creative expressions deviate from these patterns. Based on this, it is assumed that measures of the construction’s frequency, productivity and entrenchment can reversely also be used as indicators for its degree of creativity and prove that creativity in general should be considered as a gradient phenomenon.

Frequency measures and collexeme analyses of ADV-ADJ-constructions in the enTenTen15 web corpus show that highly entrenched intensification constructions use adverbs such as ‘absolutely’ or ‘very’. Patterns that are considerably less frequent contain taboo expressions coerced into the ADV-slot, such as ‘badass crazy’ or ‘batshit crazy’. These patterns have been separately analysed for frequency, productivity and entrenchment. The evidence suggests that the constructions differ substantially in their levels of productivity and entrenchment: the *batshit*-ADJ-construction, for example, is fairly unproductive as it is largely restricted to the entrenched combination with the adjective ‘crazy’. This is apparent in different frequency measures, measures of productivity (especially potential productivity (Baayen 2009)) but also collostructional analyses. But the corpus data also provide examples that are less attracted or even repelled such as hapaxes like ‘batshit original’ which suggests that these patterns may be judged to be more creative. Similar analyses will also be applied to the ADJ- *ass*-construction. The results can be compared across different intensification constructions which eventually aims for a systematic and more comprehensive rather than just exemplar-based overview of the creative potential of different adjective intensifiers.

References:

Baayen, H. 2009. Corpus linguistics in morphology: Morphological productivity. In Anke Lüdeling & Merja Kytö (eds.), Handbooks of Linguistics and Communication Science, 899–919. Berlin, New York: Mouton de Gruyter.

Goldberg, A. 2019. *Explain Me This: Creativity, Competition, and the Partial Productivity of Constructions*. Princeton UP.

Hanks, P. 2013. Creatively Exploiting Linguistic Norms. *Creativity and the Agile Mind. A Multi-Disciplinary Study of a Multi-Faceted Phenomenon*. Berlin: De Gruyter Mouton.

Perek, F. 2018. Recent change in the productivity and schematicity of the way-construction: A distributional semantic analysis. *Corpus Linguistics and Linguistic Theory* 14(1). 65–97.

Stefanowitsch, A. & S. Flach. 2017. The corpus-based perspective on entrenchment. In Hans-Jörg Schmid (ed.), *Entrenchment and the psychology of language learning: How we reorganize and adapt linguistic knowledge.*, 101–127. Washington: American Psychological Association.

Stefanowitsch, A. & S. Gries. 2003. Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8.

Larsson, Tove , Luke Plonsky and Gregory R. Hancock

Using structural equation modeling to accommodate the multivariate nature of language: The case of syntactic complexity

Despite recent advancements in statistical techniques used in corpus linguistics, there are still questions pertaining to the multivariate nature of language that our current methods cannot accommodate. In an effort to expand our analytic repertoire, this paper seeks to introduce Structural Equation Modeling (SEM) and discuss its potential for corpus linguistic analysis. Compared to traditional approaches, models in this analytical framework (e.g., measured variable path models, confirmatory factor analysis) are highly flexible in that they, for example, allow us to investigate theories involving causal effects of one or more independent variables on one or more dependent variables. Despite these and many other strengths, however, SEM remains practically unknown in corpus linguistics (though see, e.g., Gries, 2003, for an exception).

To illustrate this analytical framework, we use path models to revisit a topic that has received a fair bit of focus in recent years, namely syntactic complexity (i.e., the grammatical sophistication exhibited in language production). Specifically, we test our hypothesis based on previous research (e.g., Biber et al., 2020) that both register (academic prose vs. popular science) and disciplinary group (social sciences vs. natural sciences) can be expected to have a causal effect on three different measures of noun phrase (NP) complexity: nominal, adjectival, and prepositional modification. We use a subset of 800,000 words from the BNC and code the data using the *Tool for the Automatic Analysis of Syntactic Sophistication and Complexity* (TAASSC; Kyle 2016) to answer the following research questions:

- What is the relative importance of register and discipline on the measures of NP complexity?
- Can the hypothesized relations among the measures of complexity be explained solely by register and discipline or are there external factors at play that cause these to covary?

The results support the hypothesis stating that both discipline and register have an impact on syntactic complexity. However, unlike discipline, register has a strong effect on adjectival and prepositional modification, but only a very minor effect on nominal modification, suggesting that the nominal modification is better explained by discipline than by register. We also see that there are causal forces outside the model that exert similar influence on all three measures, showing that they do not vary independently from one another.

In this paper, we will, in an accessible way, discuss some ways in which SEM techniques can help us answer research questions that are beyond reach given commonly employed statistical methods (e.g., multiple regression). The paper will conclude with suggestions for future applications of SEM in different subdomains of corpus linguistics.

References:

Biber, Douglas, Bethany Gray, Shelley Staples & Jesse Egbert. 2020. Investigating grammatical complexity in L2 English writing research: Linguistic description versus predictive measurement. *International Journal of Academic Purposes*.

Gries, Stefan Th. 2003. Grammatical variation in English: a question of 'structure vs. function'? In Günter Rohdenburg and Britta Mondorf (eds.), *Determinants of grammatical variation in English*, 155–173. Berlin: Mouton de Gruyter.

Kyle, Kristopher. 2016. Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication (Doctoral Dissertation).

Leone, Ljubica (Lancaster University)

Linguistic changes and structural interferences: Phrasal-prepositional verbs in the Late Modern English period

The English verb system is characterized by verbs that have a phraseological constituency, namely phrasal verbs, prepositional verbs, and phrasal-prepositional verbs (Biber et al. 1999), and referred to as multi-word verbs (hereafter MWVs).

Studies conducted so far have proved that these verbs, in addition to sharing a similar internal structure, also intertwine from a diachronic perspective (Denison 1981; Hiltunen 1983; Claridge 2000). They all developed as the result of interacting processes including syntactic reanalysis, analogical generalization, and lexicalization (Denison 1981; Claridge 2000; Rodríguez-Puente 2019). It is widely accepted that phenomena of interferences characterized the development of all MWVs as there were shifts from instances from one group to another since Old English (Denison 1984, 1998; Claridge 2000). Specifically, some phrasal-prepositional verbs emerged as the result of complex processes of changes affecting extant phrasal verbs and prepositional verbs which underwent membership shifts, whereas others followed the opposite path becoming members of these latter groups (Denison 1998). These processes have been associated with early periods up to Early Modern English (Denison 1998; Claridge 2000). However, in no case have these shifts and the phenomena of interferences been examined concerning the more recent Late Modern English (LModE) period.

In order to provide evidence for the linguistic changes and membership shifts in the LModE period, the present research aims: (1) to examine the changes affecting phrasal-prepositional verbs during the years 1750-1850; (ii) to investigate phenomena of structural interferences between phrasal-prepositional verbs and other MWVs.

The present study is a corpus-based investigation undertaken on the Late Modern English-Old Bailey Corpus (LModE-OBC), a corpus covering the years 1750-1850 compiled by selecting texts from the Proceedings of the Old Bailey (https://www.oldbaileyonline.org/), London Central Criminal court.

The analysis reveals that, similarly to previous periods, the LModE time was characterized by phenomena of interferences between the various MWVs: (i) New phrasal-prepositional verbs emerged at the expense of the other two groups via analogical processes; (ii) Other phrasal-prepositional verbs underwent internal restructuring moving on into the group of phrasal verbs or prepositional verbs.

References:

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. Longman Grammar of Spoken and Written English. Harlow: Pearson Education Limited.

Claridge, Claudia. 2000. Multi-word Verbs in Early Modern English. A corpus- based Study. Amsterdam & Atlanta: Rodopi.

Denison, David. 1981. Aspects of the history of English group-verbs, with particular attention to the syntax of the ORMULUM. Oxford: University of Oxford Ph.D. Dissertation.

Denison, David. 1984. On “get it over with”. Neophilologus 68:271-277.

Denison, David. 1998. “Syntax”. In The Cambridge History of the English Language Volume IV 1776-1997, edited by Suzanne Romaine, 92-329. Cambridge: Cambridge University Press.

Hiltunen, Risto. 1983

The Proceedings of the Old Bailey Online. https://www.oldbaileyonline.org/. The Decline of the Prefixes and the Beginnings of the English Phrasal Verb: The Evidence from some Old and Early Middle English Texts. Turku: Turun Yliopisto.

Rodríguez-Puente, Paula. 2019. The English Phrasal Verb, 1650-present. History, Stylistic Drifts, and Lexicalization. Cambridge: Cambridge University Press.

Lensch, Anke

***Looking intoing showy-offy stayer-onner-for-nowers*. A corpus-based study into English morpho-syntax.**

This paper examines three Present-day English derivational schemata, which have evolved formal and functional characteristics transgressing the boundary of morphology and syntax. Today, *-er* derivatives have a wide array of meanings and the nominalizing *-er* suffix can attach to almost any word class: The suffix derives nouns from, e.g., verbs (*jog* - *jog**ger*** 'someone who Vs habitually'; *dine-din**er*** 'place where to V'), adjectives (*fresh* - *fresh**er*** 'first year university student'), prepositions (*up* - *upp**er*** 'a stimulating drug'), numerals (*five* - *fiv**er*** 'a five-pound note') and phrases, see (1). Ultimately, the only all-encompassing functional generalization befitting all *-er* nominalizations is that they denote nouns (cf. Ryder 1999: 278). English *-ing* and *-y* have evolved similarly abstract functions: *-ing* nominalizations prototypically profile processes, consider (2) (cf. Bauer et al. 2013: 207), and *-y* forms adjectives denoting a quality or characteristic of their base, consider (3) (cf. Bauer et al. 2013: 315):

(1) The resulting interwoven map of *hand-shak**er**s*, *hugg**er**s*, *kiss**er**s*, *slapp**er**s*, *slugg**er**s*, *lov**er**s* and *not-laying-a-finger-on-**er**s* should delineate the true class-structure (*The Guardian* 1993)
(2) There was much *I-told-you-so-**ing*** the following year (*The Guardian* 1990)
(3) the 'beardie-weirdie, *end-of-the-pier-**y***' associations (*The Guardian* 2004)

The derivatives in (1) to (3) illustrate that recently, the three suffixes have evolved clitic-like traits (cf. Ryder 2000: 326), as they can have scope over a whole phrase. In addition, Present-day English features highly complex derivatives involving two-fold and multiple attachment of *-er*, *-ing* and *-y*, respectively:

(4) Yesterday, though, he looked like a *stay**er**-onn**er**-for-now**er*** (*The Times* 2003)
(5) that would need some *look**ing** into**ing*** (03/09/2020 https://github.com/Pomax/Font.js/pull/71)
(6) they don't like clothes that are *show**y**-off**y*** (*The Guardian* 2000)

This repeated attachment of the suffixes is neither required syntactically nor morphologically. The quantitative and qualitative analysis of derivatives extracted from 2 billion tokens of British and American English newspaper data uncovers that although multiple marking with derivational morphology is rule-bending and still relatively rarely attested (cf. Cappelle 2010; Bauer et al. 2013: 218; Lensch 2018), it is nevertheless quite systematic.

In addition, by undertaking an analysis of the syntactic co-text of *-er* and *-ing* derivatives in historical as well as Present-day English data, this paper determines to which extent *-er* derivatives can "inherit ... the argument structure of the base verb" (Levin & Rappaport 1988: 1067), which also applies to *-ing* derivatives:

(7) *stuff**er**s* of envelopes, *knock**er**s* on doors, *organiz**er**s* of meetings (*The Guardian* 1992)
(8) baggage *check**er**s-in* and aircraft maintenanc staff
(9) Such a move would simplify the *bring**ing*** of charges.
(10) Although I do prioritise my pill-*tak**ing***, I am only human
(*The Guardian* 2003) (*The Daily Mail* 1999) (*The Guardian* 2005)

(7) to (10) show that the former direct objects of syntactically transitive verbs undergoing derivation can follow the derivatives in an *of*-phrase, or they can become the determinant of a compound (cf. Marchand 1967; Grimshaw 1990: 66; McIntyre 2015: 446). The form of some of the derivatives in this study and the analysis of their syntactic co-text offer elucidating empirical evidence showing that the derivational schemata of some English suffixes systematically blur the interface of morphology and syntax.

References

Bauer, Laurie; Lieber, Rochelle and Plag, Ingo (2013) *The Oxford Reference Guide to English Morphology*. Oxford: Oxford University Press.

Cappelle, Bert (2010) "*Doubler-upper* nouns: A Challenge for Usage-based Models of Language?". In: Onysko, Alexander & Michel, Sascha (eds.) *Cognitive Perspectives on Word-Formation*, 335-374. Berlin/New York: Mouton de Gruyter.

Grimshaw, Jane (1990) *Argument Structure*. Cambridge: MIT University Press.

Lensch, Anke (2018) "*Fixer-uppers*: Reduplication in the Derivation of Phrasal Verbs". In: Finkbeiner, Rita and Freywald, Ulrike (eds.) *Exact Repetition in Grammar and Discourse*, 158-181. Berlin: Mouton de Gruyter.

Levin, Beth and Rappaport, Malaka (1988) "Non-event *-er* Nominals: A Probe into Argument Structure". Linguistics 26: 1067-1083.

Marchand, Hans (1967) "Expansion, Transposition and Derivation". *La Linguistique* 1: 13-26. [Reprinted in Marchand 1974: 322-337].

McIntyre, Andrew (2015) "Particle-Verb Formation". In: Müller Peter, O.; Ohnheiser, Ingeborg; Olsen, Susan and Rainer, Franz (eds.) *Word-Formation: An International Handbook of the Languages of Europe*, 434-449. Berlin: de Gruyter Mouton.

Ryder, Mary Ellen (1999) "*Bankers* and *Blue-chippers*. Of *-er* Nominalizations in Present-day English". *English Language and Linguistics* 3 (2): 269-297.

Ryder, Mary Ellen (2000) "Complex *-er* Nominals: Where Grammaticalization and Lexicalization Meet?". In: Contini-Morava, Ellen and Tobin, Yishai (eds.) *Between Grammar and Lexicon*, 291-332. Amsterdam: John Benjamins Publishing Company.

Leuckert, Sven and Asya Yurchenko

The Plurality of Indian English(es): Syntactic Variation in the *Corpus of Regional Indian Newspaper English* (CORINE)

Despite a wealth of literature on Indian English (IndE) and acknowledgment of the multifaceted nature of English in India (see Lange 2017), the diversity of English(es) in the country remains critically understudied. Some of this diversity is captured in the *International Corpus of English* (ICE) as well as the Indian component of the *South Asian Varieties of English* corpus (SAVE; Bernaisch et al. 2011), but, hitherto, no corpus allows for a direct comparison of the regionally different Englishes in India. Such a corpus is a clear desideratum due to the variety of languages in contact with English as well as the varying intra- and extraterritorial forces (Labade et al. 2020) operating in different parts of the country. The newly compiled *Corpus of Regional Indian Newspaper English* (CORINE) fills this gap by giving researchers the opportunity to compare English in newspapers from most of India's states and territories.

The paper first introduces CORINE with a description of how the corpus was compiled as well as the included material in terms of word count and text types; we also comment on how interested linguists may gain access to the corpus. Then, using data from a sub-section of 5 million words of CORINE, we show regional patterns of syntactic variation in India with a case study on the 'intrusive *as*'-construction. A previous study by Koch et al. (2016) suggested that this construction (e.g. in *They call him as an idiot*) might be prevalent in written IndE, but compared its usage in a selection of South Asian and Learner Englishes and not between different Englishes in India.

The results from our case study confirm the previous finding that quotative-like verbs such as *call* and *deem* show a clear preference of 'intrusive *as*'. However, in addition, we find regional patterns that can be explained by variation in the local entrenchment of English as well as the potential impact of contact languages. Thus, in addition to introducing a new corpus to the World Englishes and corpus-linguistic research community, the paper further contributes to our understanding of the plurality of Indian English(es).

References:

Bernaisch, T., C. Koch, M. Schilk, & J. Mukherjee (2011). *Manual to the South Asian Varieties of English (SAVE) Corpus*. Giessen: Justus Liebig University, Department of English.

Koch, C., C. Lange, & S. Leuckert (2016). "'This hair-style called as "duck tail"' – the 'intrusive *as*'-construction in South Asian varieties of English and Learner Englishes." *International Journal of Learner Corpus Research* 2(2): 151-176.

Labade, S., C. Lange, & S. Leuckert (2020). "English in India: Global aspirations, local identities at the grassroots." In S. Buschfeld & A. Kautzsch (eds.), *Modelling World Englishes A Joint Approach to Postcolonial and Non-Postcolonial Varieties*. Edinburgh: Edinburgh University Press, 85-111.

Lange, C. (2017). "Indian English or Indian Englishes? Accounting for speakers' multilingual repertoires in corpora of Postcolonial Englishes." In A. Nurmi, T. Rütten, & P. Pahta (eds.), *Challenging the Myth of Monolingual Corpora*. Leiden: Brill, 16-38.

Li, Zeyu , Ulrike Gut and Philipp Meer

Patterns of co-variation in Standard Scottish English: The case of /r/, /ʍ/-/w/, and the NURSE merger

Standard Scottish English (SSE) is described as having characteristic phonological features including rhoticity, the /ʍ/-/w/ contrast, and the lack of the NURSE merger (Jones 2002). However, recent studies have found ongoing changes in SSE such as the gradual loss of rhoticity (Schützler 2013) and the /ʍ/-/w/ contrast (Schützler 2010), and a partial merger of the NURSE, EARTH and BIRD vowels (Stuart-Smith 2008). It is the aim of this paper to investigate whether there is co-variation between these trends and thus contribute to our understanding of potential phonological coherence featuring typical language variation in individual speakers. We hypothesize that a speaker with 'traditional' features will produce more tapped/trilled /r/s, have rhoticity, produce /ʍ/ more frequently, and show more distinct NURSE vowels, while a speaker with 'modern' features will have fewer or no tap/trills and /ʍ/, loss of rhoticity, and merged or merging NURSE, EARTH and BIRD vowels.

30 speakers (19f, 11m; aged 17-73) from all over Scotland giving broadcast speeches were selected from the phonologically annotated International Corpus of English (ICE) Scotland, which includes hand-corrected phonetic annotations of some automatically aligned and transcribed vocalic and consonantal segments. The following measurements were taken for each speaker: the degree of NURSE merger measured as (i) Euclidean distance between mean F1-F2 for each pair of NURSE, EARTH and BIRD and (ii) degree of centralisation as indicated by their absolute F1 and F2 *z*-scores; the realisation of /ʍ/ (rate of overall production; rate of /ʍ/ in word-initial utterance-medial and in postpausal position); and the realisation of rhotics (rate of overall tap/trill production; rate of tap/trill in onset, coda, and intervocalic position; rate of overall coda /r/ production).

We calculated cross-correlations among these variables using Spearman coefficients. Significant correlations were found between the three phonological features in SSE. The results show co-variation between the frequency of the traditional tap/trill variant in different positions and increasing Euclidean distances between NURSE-EARTH and NURSE-BIRD (Fig. 1). Moreover, the realisation of the traditional /ʍ/ variant positively correlates with the overall realisation of postvocalic /r/ (Fig. 2), the distance between EARTH-BIRD, and the centralisation of BIRD on both dimensions (Fig. 3). Our hypothesis is thus confirmed: phonological features in SSE correlate with each other, and a 'traditional' Scottish English user of one phonological variant tends to produce other traditional variants as well.

References:

Jones, Charles. (2002). *The English language in Scotland: an introduction to Scots*. Tuckwell Press.

Schützler, Ole. (2013). "The Sociophonology and Sociophonetics of Standard Scottish English (r)". In Peter Auer, Javier Caro, & Götz Kaufmann (eds.), *Language Variation - European Perspectives IV*. Amsterdam: Benjamins, 215-228.

Stuart-Smith, Jane. (2008). Scottish English: phonology. In Bernd Kortmann & Clive Upton (eds.), *Varieties of English. The British Isles*. Berlin: Mouton, 48-70.

Schützler, Ole. (2010). Variable Scottish English Consonants: the Cases of /ʍ/ and non- prevocalic /r/. *Research in Language* 8, 5-21.

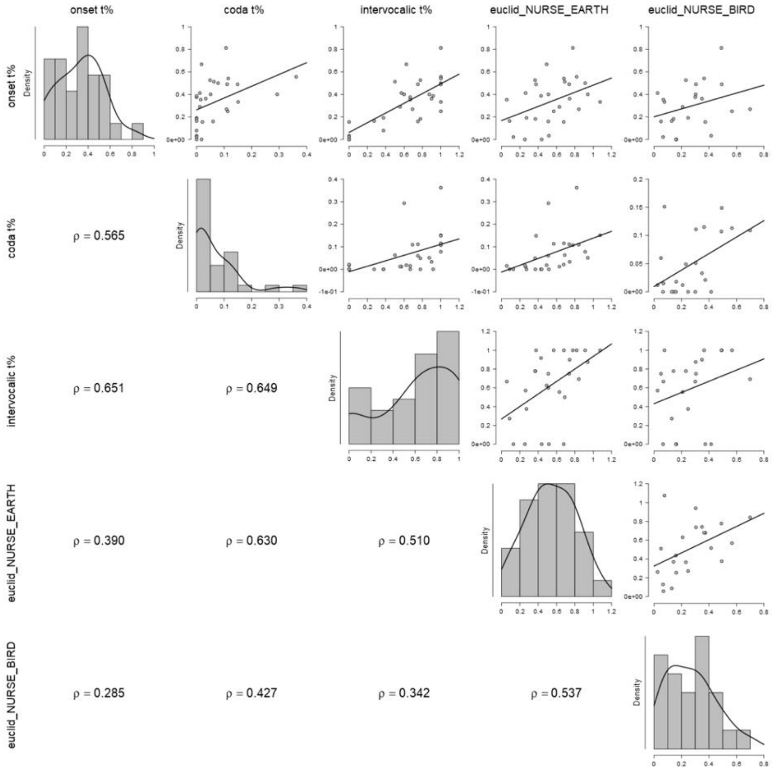

Figure 1. Correlations between the realisation of rhotics and NURSE merger.

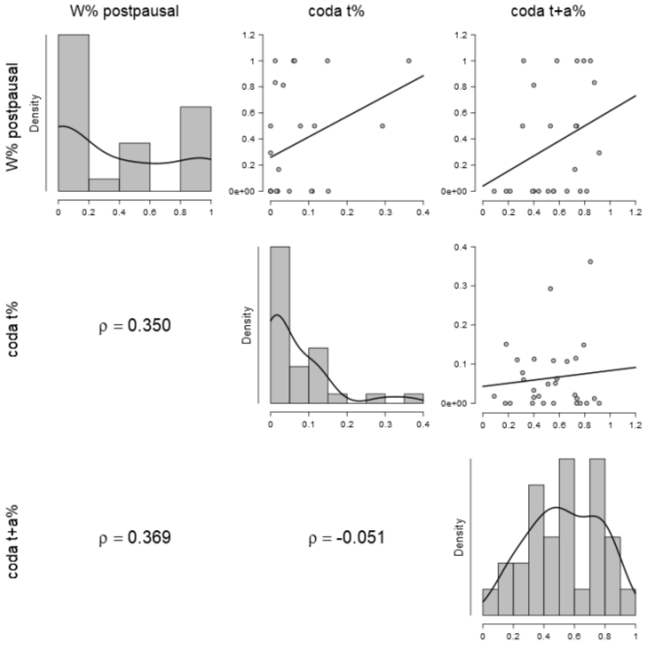

Figure 2. Correlations between the realisation of /ʍ/ and rhotics.

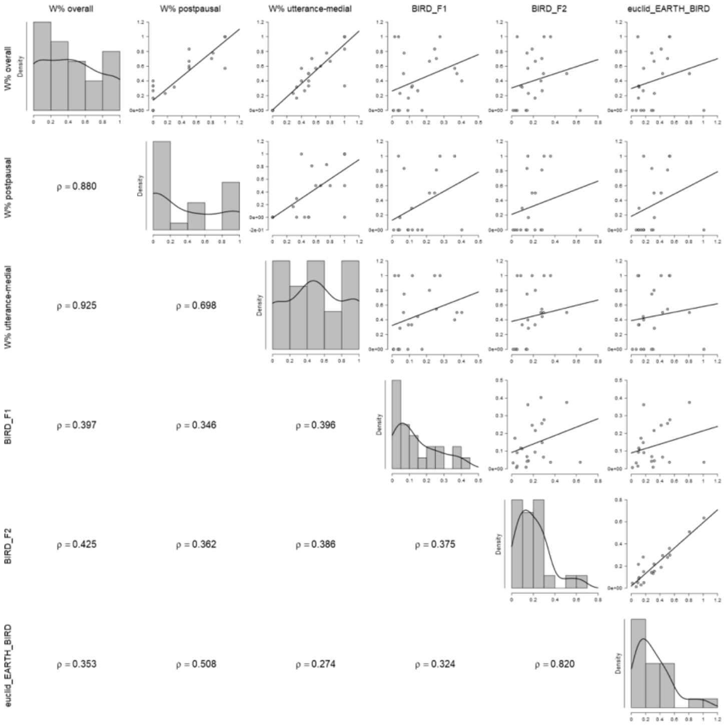

Figure 3. Correlations between the realisation of /ʍ/ and NURSE merger.

Mahler, Hanna

The interpretative progressive: A case of grammatical constructionalisation?

The interpretative progressive (e.g. "when he says 'under my administration' he's *implying* that it's because of policies that he's enacted"), one of the 'non-central' uses of the English progressive, has so far received only little scholarly attention. Out of the studies that consider it, only a few use corpora (of mainly written language) to support their argumentation (e.g. Kranich 2010, Smith 2005, Leech et al. 2009).

While most scholars agree that the increasing frequency of the interpretative progressive contributed to the overall increase in the use of the progressive in the 20th century, opinions differ regarding the importance of this usage (e.g. Kranich 2010:223, Leech et al. 2009:136). The answer seems to be related to register, as the interpretative progressive is commonly classified as a feature of oral language – stylistic changes such as colloquialisation therefore also play a role.

This paper presents the first corpus-linguistic analysis of the interpretative progressive based on spoken data only. Using the framework of constructionalisation (Traugott & Trousdale 2013), the study investigates whether the development of the interpretative progressive can be regarded as a case of grammatical constructionalisation. Instances of the interpretative progressive with verbs of communication are retrieved from the spoken section of the Corpus of Contemporary American English (Davies 2008) and are analysed regarding the criteria for constructionalisation: increasing productivity, increasing schematicity, and decreasing compositionality (Traugott & Trousdale 2013:22).

While the interpretative progressive is the only use of the progressive requiring two clauses (which supports its status as an independent construction), these two components can be syntactically distant, and the interpreted clause can also be omitted. Analysis of a preliminary sample of 186 data points shows no "fixing of form" over the span of the corpus with regard to the presence or absence of introductory clauses or lexical cues. Furthermore, there is no apparent decrease in the variability concerning the placement and the aspectual choice of the interpreted clause. As there is also no increase in type and token frequency, increasing productivity also seems dubitable.

The preliminary sample from the COCA does therefore not provide sufficient evidence to support the theory that the interpretative progressive is subject to grammatical constructionalisation. This has wider implications for the conceptualisation of the different uses of the progressive construction and the debate around a "basic meaning" (Kranich 2010:72-76).

References:

Davies, Mark. 2008-. *The Corpus of Contemporary American English (COCA): 560 Million Words, 1990-Present*. www.english-corpora.org/coca/.

Kranich, Svenja. 2010. *The Progressive in Modern English: A Corpus-Based Study of Grammaticalization and Related Changes* (Language and Computers: Studies in Practical Linguistics 72). Amsterdam: Rodopi.

Leech, Geoffrey, Marianne Hundt, Christian Mair & Nicholas Smith. 2009. *Change in Contemporary English: A Grammatical Study* (Studies in English Language). Cambridge: Cambridge University Press.

Smith, Nicholas. 2005. *A Corpus-Based Investigation of Recent Change in the Use of the Progressive in British English*. Lancaster: Lancaster University Dissertation.

Traugott, Elizabeth & Graeme Trousdale. 2013. *Constructionalization and Constructional Changes* (Oxford Studies in Diachronic and Historical Linguistics). Oxford: Oxford University Press.

Matsumoto, Noriko

The Ongoing Change from the *Remain-UnVed* Sequence to the *Go-UnVed* Sequence

This paper shows the ongoing change from the *remain-unVed* sequence to the *go- unVed* sequence, relying on the COHA and Coronavirus Corpus. The historical development of the *go-unVed* sequence is discussed in comparison with the uses of the *remain-unVed* sequence, which are significantly related to the *go-unVed* sequence, as in (1)-(2).

(1) His remarks went unnoticed by his staff.
(2) His remarks remained unnoticed by his staff.

From a semantic standpoint, the verb *go*'s nearest functional equivalent would be the verb *remain*, because both the *go-unVed* and *remain-unVed* sequences convey a sense of continuation. *Go* in (1) functions as a marker of evaluative modality, which signals the modal notion of counter-normativity. Bourdin (2003) points out that evaluative modality with respect to the *go-unVed* sequence shows an impersonal quality, which is equivalent to a speaker's negative judgment on behalf of society. However, *remain* in (2) does not involve such a negative judgment. The *remain-unVed* sequence offers an objective report. In Figure 1, the COHA shows that the *go-unVed* sequences have been recently gaining in currency in that they are replacing the *remain-unVed* sequences. This paper proposes a working hypothesis: the *go-unVed* sequence represents ongoing historical development that results in significant shifts in frequency of the *go-unVed* sequence. Based on our corpus data, this paper also shows how the ongoing change of the *go-unVed* sequence is closely related to inflectional categories of the verb *go*, adjective selection, and historical development in a complicated way.

There are four main findings from our corpus data. First, in Figures 2-4, the COHA shows that there are three types of the ongoing change from the *remain-unVed* sequence to the *go-unVed* sequence, the replacement type, the predominance type where *go-unVed* sequences outnumber *remain-unVed* sequences, and the indeterminacy type. Second, in Figures 5-7, the Coronavirus Corpus shows that there are three patterns of the relationship between the *go-unVed* and *remain-unVed* sequences, the superiority pattern where *go- unVed* sequences outnumber *remain-unVed* sequences, the inferiority pattern where *remain-unVed* sequences outnumber *go-unVed* sequences, and the ups-and-downs pattern. Third, both the COHA and Coronavirus Corpus show the roughly similar distribution of the top ten adjectives used most frequently. Fourth, concerning inflectional categories of the verb *go*, both the COHA and Coronavirus Corpus show that present forms are predominant. The *go-unanswered* and *go-unheeded* sequences are atypical.

From these findings, it is fair to state that the replacement type is related to the superiority pattern, and that the indeterminacy type is related to the ups-and-downs pattern. The *go-unnoticed* sequence is representative of the former, and the *go- unanswered* sequence the latter. It is clear that the ongoing change from the *remain-unVed* sequence to the *go-unVed* sequence is observed. It can be concluded that the verb *go* as a marker of evaluative modality plays a key role in the ongoing change.

Reference:

Bourdin, P. 2003. On two distinct uses of *go* as a conjoined marker of evaluative modality. In R. Facchinetti, M. Krug, & F. Palmer, eds., *Modality in Contemporary English*, 103-128. Berlin/New York: Mouton de Gruyter.

Milička, Jiři and Denisa Šebestová

Human Friendly Corpus Query Language

This study addresses the pressing question how to make corpora accessible to users who are not able or willing to learn a programing language, or who are not granted full access to the data. The design of query methods and user interface is not a mere technical problem, as it shapes the whole paradigm of corpus linguistics, while being shaped by the same paradigm in return. The more specific a tool is the more tightly it is bound to the given paradigm, the less inventiveness it allows or at least encourages.

Corpus Query Language (CQL), as used in SketchEngine, Czech National Corpus web applications and many other services, is a popular Swiss army knife among formal query languages, allowing for great variability while being expressive and unambiguous.

The unambiguity and strict formal definition is a requirement to be a fluent tool for communication with a query processing engines. Nevertheless, for many researchers, the strictness of formal languages such as CQL poses a major obstacle. Not to mention its notation, which discriminates against dyslexic users; the whole idea of formal language, which generates meaning bottom-up by combining specific building blocks according to strict rules, is very foreign not only to untrained hobbyists but even to many professionals.

In an attempt to address this issue, we have developed an automatic translator from a near-natural language to CQL. It allows for submitting queries such as "find adjectives followed by the lemma 'book'" or "I want you to search for all feminine substantives syntactically dependent on a first person verb" or "find a contiguous sequence of at least five nouns, all starting with 'p'". The translator does not require any strict syntactical patterns (unlike typical formal languages), even allowing for a certain amount of typing errors, using the redundancy offered by natural language. The translator does not parse the query syntactically nor logically; instead it unpacks meaning in a discriminative way: it builds conjectures about what the user wants to communicate and then searches for parts of the query that confirm, disprove of modify the original conjectures. This method enables the translator to process even complicated or fragmentary queries.

The users still need to follow certain rules. For instance, the queried words must be entered in quotations marks; the query needs to be linear, i.e. no referring back to previous parts of the query. Moreover, they need to know the linguistic metalanguage and the architecture of the corpus, such as lemmatization or morphological tagging, and be aware of the possibilities offered by the query engine and the dataset.

At the time of abstract submission, a working Czech–CQL translator is available online (http://alpha.korpus.cz). We plan to have generalized our solution to English by the ICAME conference. This should be relatively uncomplicated as the English morphology is less complex than the Czech one.

Moessner, Lilo

Modality and the English subjunctive. A diachronic study

It is usually taken for granted that at least in British English the subjunctive alternated with modal constructions in all historical periods and that in Present-Day English (PDE) the former was largely replaced by the latter. The representatives of this view argue that the frequency decrease of the subjunctive is a consequence of the shrinking verbal paradigm starting at the end of the Old English period. In their studies the subjunctive is defined as one of the realisations of the grammatical category mood, the others being indicative and imperative. This definition of the subjunctive has been rejected for PDE in more recent publications by Huddleston and Pullum (2002: 77) and Bas Aarts (2012: 1), who define the subjunctive as a clause type, and by Ruohonen (2017: 23), who pleads for the inclusion of its meaning component in the definition of the subjunctive.

I will argue along Ruohonen's lines that a definition of the subjunctive as a combination of form and meaning is a promising starting-point for a more appropriate description of the frequency development of the English subjunctive. I will accept the traditional definition of the subjunctive as a realisation of the grammatical category mood, but add the semantic feature that it expresses root modality. Modal constructions, by contrast, can express root modality or epistemic modality (James 1986: 13). Two hypotheses which will be tested in my paper follow from this constellation: a) the balance between subjunctives and modal constructions will change in favour of the subjunctive, when their frequency is compared only to that of those modal constructions which express the same (= root) modality, b) some modal constructions contribute more to the change of this balance than others.

Since in mandative constructions the suasive matrix element is the strongest predictor for the choice between subjunctives and modal constructions (Hundt 2019: 231), I will analyse main/independent clauses where this factor is not involved. It is my aim to establish the strength of the influence of the modality expressed by the relevant verbal syntagms in the historical periods Old English, Middle English, and Early Modern English. My data will come from the respective parts of the Helsinki Corpus of English Texts.

References:

Aarts, Bas (2012), 'The subjunctive conundrum in English', Folia Linguistica 46/1, pp. 1-20. Huddleston, Rodney and Geoffrey Keith Pullum (2002), The Cambridge Grammar of the English Language, Cambridge: Cambridge University Press.

Hundt, Marianne (2019), 'It is important that Mandatives (should) be studied across different World English and from a Construction Grammar Perspective', in Paloma Núñez Perteja, María López Couso, Belén Mendez Nay and Ignacio Palacios Mantinez (eds), Crossing Linguistic Boundaries: Systemic, Synchronic and Diachronic Variation in English, London: Bloomsbury, pp. 211-238.

James, Francis (1986), Semantics of the English Subjunctive, Vancouver: University of British Columbia Press.

Ruohonen, Juho (2017), 'Mandative Sentences in British English: Diachronic Developments in Newswriting between the 1990s and the 2010s', Neuphilologische Mitteilungen 118/1, pp. 171-200)

Pitzle, Marie-Luise and Ruth Osimk-Teasdale

Introducing VOICE Online 3.0 – A new web tool for research on spoken ELF interactions

Since the first online release of the Vienna-Oxford International Corpus of English (VOICE) in 2009, and additional releases in subsequent years (i.e. VOICE XML in 2011, VOICE 2.0 Online, VOICE POS XML and VOICE POS Online in 2013), thousands of users world-wide have used the online corpus both, for linguistic research and for university teaching on the nature of spoken English as a lingua franca (ELF). VOICE Online 1.0 and 2.0 already included features like different layout and annotation styles (voice, plain, kwic), flexible display of individual mark-up features (i.e. an on/off function), filters and bookmarks. In 2013, a separate web interface was launched for the part-of-speech (POS) tagged version of VOICE (VOICE POS Online). Already then, the implementation of innovative web features (such as the flexible display of mark-up) was made possible by the detailed nature of conversational transcripts in VOICE and the accompanying high technical standards. VOICE transcripts were converted to fully TEI-conform XML texts that served as the backbone for the original VOICE Online interfaces (released in 2009 and 2013).

Building on and expanding these standards, the VOICE CLARIAH project was launched in 2020 as a cooperation between the Austrian Centre for Digital Humanities and Cultural Heritage (Austrian Academy of Sciences) and the University of Vienna. The aim of the project is to ensure continued and stable open-access to VOICE - not only as a downloadable corpus, but also as an enhanced web tool for research on spoken ELF interactions.

In this software demonstration, we will introduce the new VOICE Online 3.0 interface developed in the VOICE CLARIAH project and scheduled for release in summer 2021. While VOICE Online 3.0 includes all functions of the initial VOICE web interfaces, it offers a number of new and increased web functions. These include a range of additional filter, style and search functions that will increase the usability of the online corpus for qualitative and quantitative work. One novel core feature will be the searchability of select mark-up features for spoken language (pauses, words spoken simultaneously) and multilingual data (L1, LN, LQ). Moreover, for the first time, the regular annotated VOICE interface is merged with the VOICE POS Online interface. This means that lexical searches, searches for conversational mark-up and annotated POS and lemma tags will be possible in the same web interface - and can therefore be intuitively combined also by non-specialist corpus users.

Our presentation will introduce these new key features of VOICE Online 3.0. It will also highlight their increased research potential for the study of ELF interaction and for research on spoken language more generally. In this spirit, we will provide some background also on the technical implementation of VOICE Online 3.0 and discuss some key components of the new frontend and backend infrastructure. These include the creation of a new ‘merged’ and tokenized version of the VOICE XML texts (incorporating POS and conversational mark-up in one TEI-compatible XML file), the integration of current web technologies (like node.js, json) and NoSketch Engine.

Puga, Karin, Kathrin Kircili and Sandra Götz-Lehmann

Syntactic segmentation of spoken corpus data: What prosody can contribute

To analyze the clause structure of spoken language, Biber et al. (1999) suggest the segmentation into clausal units, "consisting of an independent clause together with any dependent clauses embedded within it" (ibid: 1069), and non-clausal units, i.e. words, phrases or unembedded dependent clauses which may enter into relations with other elements to form a clausal unit. Most corpus-based syntactic segmentation schemes, however, rely on the transcriptions alone (cf. Biber et al. 1999; Szaszak et al. 2011; Wagner 2015), which, can lead to segmentation difficulties, especially when analyzing spontaneous conversations: For instance, it can be ambiguous to decide whether inserts (e.g. discourse markers) should be separated from a following C-unit or not, as they could either stand on their own and thus be segmented as non-clausal units or they could join with a following unit to form a larger structure (e.g. [oh|yeah|that's pretty] vs. [oh|yeah that's pretty] vs. [oh yeah that's pretty]). Following Biber et al. (1999: 1076), "[t]he only criterion which helps segmentation here is whether the insert is separated prosodically (in terms of intonation) from the following unit".

Against this background, our proposed paper takes into account previous suggestions in describing the prosody-syntax interface (e.g. Du Bois et al. 1992; Selting 2000) and suggests an approach to syntactic segmentation that combines syntactic-based approaches (Biber et al 1999) with those that take a prosodic focus (Beckman and Pierrehumbert 1986). The study describes the exact syntactic contexts in which syntactic segmentation needs to be complemented by prosodic analyses to become fully accurate. In doing so, we will analyze the Louvain Corpus of Native English Conversation where each unit is annotated for various variables, including the LENGTHOFC-UNIT vs. LENGTHOFIU, NUMBEROFINSERTS, etc. In our analysis, we fit regression models to predict the factors significantly increasing the likelihood of a purely syntactic segmentation to become more/less accurate. While the results of a first pilot study indicate a considerable overlap of clause- and ip-boundaries, they also indicate factors that contribute strongly to an increase/decrease of ambiguities in segmentation (e.g. C-units that have multiple clauses but only one f0-contour or vice versa). We will discuss our findings in the light of their benefits on conducting syntactic segmentation of spoken corpus data.

References:

Beckman, M. and J. Pierrehumbert (1986): "Intonational structure in Japanese and English", Phonology Yearbook 3, 255-309.

Du Bois, J. W., Schuetze-Coburn, S., Paolino, D., and Cummings, S. (1992): "Discourse transcription", Santa Barbara Papers in Linguistics 4, Dept. of Linguistics, University of California, Santa Barbara.

Selting, M. (2000): "The construction of units in conversational talk". Language in Society 29, 477-517.

Szaszák, G., K. Nagy and A. Beke (2011): "Analysing the correspondence between automatic prosodic segmentation and syntactic structure", Interspeech 2011, 1057-1060.

Wagner, M. (2015): "Phonological Evidence in Syntax", Syntax—Theory and Analysis, ed. T. Kiss and A. Alexiadou. Berlin: Mouton de Gruyter, 1154-1198.

Biber, D. S. Johansson, G. Leech, S. Conrad and E. Finegan (1999): Longman
Grammar of Spoken and Written English. Harlow: Pearson.

Ranaweera, Mahishi

A corpus-based study of pragmatic markers in the spoken discourse of women in Sri Lankan English

Pragmatic markers have been widely researched. Some research concludes that the use of pragmatic markers (PMs) depends on the function and purpose of communication rather than the demographic details of the speaker (Aijmer, 2002). However, other research notes that certain pragmatic markers are associated with sociolinguistic variables such as the age and gender of a speaker (Stenström, 2014). Recent research on English pragmatic markers also suggests that we need to examine their occurrence across varieties of English to observe possible variations (Aijmer, 2013). There is research on many aspects of pragmatic markers in varieties of Englishes such as British English, Australian English, Swedish English etc. (Williams, Mulder, Moore, 2017). However, research on the function and use of pragmatic markers with regard to lesser-known Englishes such as Sri Lankan English is rare.

This study examines the use and the functions of English pragmatic markers by female artists and female entrepreneurs in semi formal interviews in Sri Lanka. The items of PMs considered in this study are any words which are optional than obligatory in an utterance and are difficult to be categorized into any traditional word class (Andersen, 2001). Example items would be 'you know', 'I mean', 'like', 'well' etc. The study answers the following research questions: What are the pragmatic markers used by female artists and entrepreneurs in Sri Lanka, and what are their functions? What pragmatic markers are specific to Sri Lankan English (SLE)?

The data is taken from a purpose- built corpus, and is analysed both quantitatively and qualitatively. The corpus consists of twenty online videos featuring semi-formal interviews. Each interview is between an anchorman/anchorwoman and a female artist or an entrepreneur. There are equal numbers of interviews from both categories. Only the data from the interviewee was selected for analysis. The interviews were transcribed and a sample of 2,000 words from each interviewee was selected for inclusion in the corpus. The total word count of the corpus is 40,000. This was then manually analysed to identify instances of pragmatic markers and their context of occurrence, to assist in the classification of their functions. The frequency of the pragmatics markers was analysed to comment on the preference of use. Preliminary results indicate that while most of the pragmatic markers used by Sri Lankan English speakers are common to BrE and IndE, there is also evidence of pragmatic markers native to Sri Lankan English.

References:

Aijemr, Karin. 2002. English discourse particles: evidence from a corpus. Amsterdam and Philadelphia: John Benjamins.

Aijmer, Karin. 2013. Understanding pragmatic markers: A variational pragmatic approach. Edinburgh: Edinburgh University Press.

Andersen, Gisle. 2001. Pragmatic Markers and Sociolinguistic Variation. Amsterdam/Philadelphia: John Benjamins.

Stenström, Anna-Brita. 2014. Teenage Talk from General Characteristics to the Use of Pragmatic Markers in a Contrastive Perspective. London: Palgrave Macmillan UK.

Williams, Cara P., Jean Mulder & Erin Moore. 2017. The pragmatic functions of sort of in Australian English. Presented at the Conference of the Australian Linguistic Society.

Ren, Haoshan

Abstract nouns in Chinese-learner English argumentative writing

Abstract nouns and their patterns of use are highly distinctive and context-specific (Laso, & John, 2013). They are shown to be a difficult target for L2 learners (Lee & Spinner, 2012). However, abstract nouns have only been studied for their discoursal functions (Flowerdew, 2010; Schmid, 2012) with little focus on their structural and collocational patterns. Therefore, this study used a rigorous corpus-based methodology to investigate 1) the frequent collocational and colligational patterns of high-frequency abstract nouns; 2) the prototypical verbs and verb-constructions co-occurring with high-frequency abstract nouns. Specifically, this study focuses on writings produced by advanced L1 Chinese learners of English.

Frequent nouns used in argumentative writings were first retrieved from the Chinese sub-corpus of the International Corpus of Learner English (C-ICLE) using TagAnt and AntConc (Anthony, 2016, 2019). This list was then manually matched to the list of abstract nouns retrieved from the MRC database (Coltheart, 1981) with a low concreteness score. Among the resulting 81 high-frequency abstract nouns, 6 abstract nouns with frequency greater than 130 in C-ICLE were selected for the current analysis. Concordance lines containing both singular and plural forms of each abstract noun were then manually analyzed to identify verb constructions with abstract nouns (Goldberg, 1995, 2006). These results were then compared to the same nouns and patterns in MICUSP argumentative essays.

Findings showed various and distinctive collocational and colligational tendencies of each abstract noun used by advanced Chinese L2 writers, with a few abnormal high-frequency constructions (e.g. are lack of). A few verb constructions were shown to be frequently used with all six abstract nouns regardless of their semantic differences. These findings reveal that the advanced Chinese learners of English conform to specific choices of conceptual mappings to constructions typical to abstract nouns. These mappings do not align altogether with patterns found in argumentative essays in MICUSP.

References:

Anthony, L. (2016). TagAnt (Version 1.2.0) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/software

Anthony, L. (2019). AntConc (Version 3.5.8) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/software

Coltheart, M. (1981). The MRC psycholinguistic database. The Quarterly Journal of Experimental Psychology Section A, 33(4), 497-505.

Flowerdew, J. (2010). Use of signalling nouns across L1 and L2 writer corpora. International Journal of Corpus Linguistics, 15(1), 36-55.

Goldberg, A. E. (1995) Constructions: A construction grammar approach to argument structure. Chicago: University of Chicago Press.

Goldberg, A. E. (2006) Constructions at work: The nature of generalization in language. Oxford: Oxford University Press.

Laso, N. J., & John, S. (2013). A corpus-based analysis of the collocational patterning of adjectives with abstract nouns in medical English. I. Verdaguer, NJ Laso & D. Salazar, Biomedical English: A Corpus-Based Approach, 55-71.

Lee Amuzie, G., & Spinner, P. (2012). Korean EFL learners' indefinite article use with four types of abstract nouns. Applied Linguistics, 34(4), 415-434.

Schmid, H. J. (2012). English abstract nouns as conceptual shells: From corpus to cognition (Vol. 34). Walter de Gruyter.

Ren, Haoshan and Viviana Cortes

Functional language in academic lectures – an investigation of corpus linguistic findings nad implication for language assessment

Academic lectures have been one of the most studied spoken academic genres, especially from a corpus linguistics perspective. These corpus studies demystify academic discourses for the non-native English speaking participants in English-dominant academic discourses, including international students taking English medium classes and non-native speaking instructors who are teaching in English. However, existent literature covers a wide variety of linguistic units studied under different functional frameworks, which complicate the process of applying current findings to future research, assessment, or pedagogical uses. Besides, although corpus-informed language features start to emerge as a construct in language testing of non-native speaking instructors (e.g., Römer, 2017) in addition to the traditionally tested constructs (e.g., pronunciation), no empirical evidence has shown its relationship with conventionally used constructs such as pronunciation. This issue has been increasingly pertinent to the population of international teaching assistants (ITAs), whose academic trajectories are dependent on ITA placement tests that assess their ability to use authentic language for academic lectures.

Therefore, this study first presents a systematic review of 26 corpus studies on academic lectures in the past decade, revealing three main types of linguistic units as well as two types of functional frameworks employed in these analyses. Specifically, linguistic units analyzed in recent literature include frequency-based units (mainly lexical bundles), function-driven units (questions, importance/relevance markers, personal metalinguistic markers, and pragmatic force modifiers), and structural units (single words and grammatical structures). This review not only provides a synthesized and practical framework for the functional language used in academic lectures, but also serves as a foundation to the following empirical study, which validates the use of frequency-based functional units as an assessment construct in evaluating international teaching assistants' test performances, especially in tasks where they perform a teaching demonstration of an academic lecture.

For the empirical study, a linear regression analysis was conducted on data collected from 30 ITA teaching demonstrations in an ITA test to model the relationship between pronunciation, functional language use, and overall teaching effectiveness. Pronunciation and overall teaching effectiveness were rated by seven expert raters using a connected design (Myford & Wolfe, 2000). Audio recordings were transcribed verbatim into texts for corpus analysis of frequency-based functional language using the list provided in Liu and Chen (2020). The frequency data were processed using a python code. Results show that both pronunciation and functional language use are significant predictors of ITAs' overall teaching effectiveness, thus providing evidence for establishing functional language use as a potential construct for ITA assessments. This study supports the connection between corpus linguistic studies and their applications in language assessment.

References:

Römer, U. (2017). Language assessment and the inseparability of lexis and grammar: Focus on the construct of speaking. Language Testing, 34(4), 477–492.

Liu, C. Y., & Chen, H. J. H. (2020). Analyzing the functions of lexical bundles in undergraduate academic lectures for pedagogical use. English for Specific Purposes, 58, 122-137.

Myford, C. M., & Wolfe, E. W. (2000). Strengthening the ties that bind: Improving the linking network in sparsely connected rating designs. ETS Research Report Series, 2000(1), i-34.

Romasanta, Raquel P.

Geographical location as predictor of language variation in postcolonial Englishes

This paper explores the role of geographical location in the inter-varietal variation detected in the verbal complementation system of postcolonial Englishes. The point of departure is the fundamental principle in dialectology that "geographical proximity between dialects should predict dialectal similarity between dialects" (Szmrecsanyi 2012: 837). In other words, geographically close English varieties are prone to show more similarities in their complementation systems than geographically distant ones. This was noted as early as 1980 in Streven's World Map of English Model, where it is stated that each form of English "normally exhibits similarities with other forms of English in the same geographical area" (Strevens 1980: 85). In order to test this principle, the focus in this presentation will be on the non-categorical variability of the clausal complementation of the verb REGRET (*she regrets selling her car* vs *she regrets that she sold her car*) in African varieties of English, namely South African, Nigerian, Ghanaian, Kenyan, and Tanzanian Englishes, and the two main Inner Circle varieties, included here for comparison (British and American English). According to this principle, we could expect to see a divide between East, West, and Southern African English varieties, as well as differences between African varieties and those of the Inner Circle.

The sample includes 2,089 instances of *regret* + (S) *-ing* clause and *regret* + *that*-clause retrieved from the GloWbE corpus. In order to identify the determinants of complement variation, I analyzed the intra-linguistic factors frequently adduced to play a role here: meaning of the main and subordinate verbs, animacy of the subject, voice of the complement clause, presence of negative markers in the complement clause, and intervening material between the two clauses, among others. Multidimensional aggregational analyses of two types, hierarchical cluster analysis and non-hierarchical phylogenetic networks (NeighborNet), have allowed me to visually aggregate similarities and distances between the varieties. Preliminary results seem to confirm that geographical location predicts variation to a fairly large extent, in that three groups can be differentiated: (i) East Africa, with Kenyan and Tanzanian Englishes; (ii) West Africa, with Nigerian and Ghanaian Englishes; and (iii) a more heterogeneous group including British, American, and South African Englishes. What draws British and American English together is not location but probably their status as Inner Circle varieties; the linguistic proximity between South African English and Inner Circle varieties might be partially interpreted in terms of phase of development according to Schneider's (2007) Dynamic Model, since South African English is in phase 4 while the other African varieties are in phase 3. This indicates that, as expected, other factors intersect with location, as is the case with Inner vs. Outer Circle status and phase of development, among others.

References:

Schneider, Edgar W. (2007). *Postcolonial English: Varieties around the world.* Cambridge: Cambridge University Press.

Strevens, P. (1980). *Teaching English as an international language: From practice to principle*. Oxford: Pergamon Press.

Szmrecsanyi, B. (2012). Typological profile: L1 varieties. In B. Kortmann and H. Paulasto (Eds.), *The Mouton world atlas of variation in English* (p. 826-843). Berlin: Mouton de Gruyter.

Rodríguez-Puente, Paula

Suffix competition across registers: On the development of *-ity* and *-ness* in the Late Modern English period

This paper builds on previous research which sought to trace the development of two deadjectival nominalizing suffixes, the Romance *-ity* and the native *-ness*, during the Early Modern English period. Rodríguez-Puente (2020) compared their distribution across seventeen registers representative of the speech-written and formal-informal continua, demonstrating that *-ness* decreased in favour of -*ity* between the sixteenth and the early eighteenth centuries, a change which seems to have begun in formal written registers and spread towards speech-related ones, probably aided by a general trend towards the adoption of a more literate style particularly during the eighteenth century (Biber & Finegan 1997). In this paper I seek to explore the later history (1700- 1900) of the two suffixes in sixteen different registers obtained from four corpora: *A Representative Corpus of Historical English Registers* 3.2 (ARCHER), the *Penn Parsed Corpus of Modern British English* (PPCMBE), the *Corpus of Historical English Law Reports, 1535-1999* (CHELAR) and the *Old Bailey Corpus* (OBC).

Despite the importance of register analysis in the development of languages (Biber & Gray 2013), few investigations have explored the interplay between suffix usage and register during the Late Modern English period. Cowie (1998) is, to the best of my knowledge, the only study which examines the two suffixes in the various registers of ARCHER from 1650 to 1990 based on a measure of aggregation of new types over time, though she concludes that register is not determinant in their use. In this paper I seek to verify Cowie's results with a more fine-grained analysis focusing exclusively on the British variety and including a wider range of registers which are examined in terms of two measures: type counts and the introduction of new types over time. Preliminary results suggest that the learned, Romance suffix *-ity* continued as the predominant suffix during the eighteenth and early nineteenth centuries, except in registers representative of the spoken, informal variety, i.e. trial proceedings. In turn, with the progressive democratization of language and the colloquialization of written registers (see, e.g., Hundt & Mair 1999; Hiltunen & Loureiro-Porto 2020), *-ness* begins to gain ground towards the end of the nineteenth century.

References:

Biber, Douglas & Edward Finegan. 1997. Diachronic relations among speech-based and written registers in English. In Terttu Nevalainen & Leehna Kahlas-Tarkka (eds.), *To Explain the Present: Studies in the Changing English Language in Honour of Matti Rissanen*, 253-275. Helsinki: Société Néophilologique.

Biber, D. & Gray, B. 2013. Being specific about historical change: The influence of sub- register. *Journal of English Linguistics* 41(2): 104-134.

Cowie, Claire. 1998. *Diachronic Word-formation: A Corpus-based Study of Derived Nominalizations in the History of English*. Doctoral dissertation, University of Cambridge, United Kingdom.

Hiltunen, Turo & Lucía Loureiro-Porto. 2020. Democratization of Englishes: Synchronic and diachronic approaches. *Language Sciences* 79. https://doi.org/10.1016/j.langsci.2020.101275

Hundt, Marianne & Christian Mair. 1999. "Agile" and "uptight" genres: The corpus- based approach to language change in progress. *International Journal of Corpus Linguistics* 4(2): 221–242.

Rodríguez-Puente, Paula. 2020. Register variation in word-formation processes: The development of *-ity* and *-ness* in Early Modern English. *International Journal of English Studies* 20(2): 147-169.

Römer, Ute

"I can't sing and dance but I can cook": How do constructions with modal verbs develop in second language learners of English?

Recent research in corpus- and usage-based linguistics has indicated that we acquire language by learning constructions. This applies to both first language (L1) learners (Goldberg, Casenhiser & Sethuraman, 2004; Tomasello, 2003) as well as second language (L2) learners (Ellis, 2002; Ellis & Cadierno, 2009; Author et al., 2016). Constructions have been described as the building blocks of language, and defined as conventionalized pairings of form and meaning that are entrenched in the speaker's mind (Bybee, 2010; Goldberg, 1995). Compared to L1 acquisition, our understanding of the development of constructions in L2 English learners is still rather limited. Existing corpus-based studies on L2 construction development have relied almost exclusively on data produced by small numbers of learners, and have focused on small sets of constructions.

Applying a combination of methods from Corpus Linguistics and Natural Language Processing, the study reported on in this paper uses data from a large corpus of writing produced by L1 German and L1 Spanish learners of English to investigate how knowledge of a large set of verb-argument constructions (VACs) develops in L2 English learners across proficiency levels from low beginner to upper intermediate (CEFR levels A1 to B2). The focus of the paper is on a subset of VACs that contain modal auxiliaries in combination with lexical verbs. The specific research questions addressed in this paper are: (1) How does the distribution of verbs in VACs with modals in learner production develop across proficiency levels; and (2) Are there significant observable differences in the acquisition of verbs in VACs with modals between L1 German and L1 Spanish learners of English?

To address these questions, data on selected high-frequency VACs containing modal verbs, including SUBJ-MODAL-V (e.g., I can cook) or SUBJ-MODAL-V-DOBJ (e.g., we will speak English), was exhaustively extracted from a 6-million word subset of EFCAMDAT, the Education First-Cambridge Open Language Database (Geertzen et al., 2013). The texts in the EFCAMDAT subset were divided by learner level and L1 into eight subcorpora (e.g., German-A1, Spanish-B2). Using a customized Python script, we generated frequency-sorted lists of verbs in VACs for each level and L1 which allowed for comparisons of learners' verb selection preferences across levels/L1s. Our paper will discuss findings on observed changes in modal VAC productivity and complexity from low to higher proficiency levels, and on similarities and differences in the verb selections of learners from different L1 backgrounds. These findings help us expand our understanding of the processes that underlie L2 construction acquisition.

References:

Author et al. 2016.

Bybee, J. (2010). Language, Usage and Cognition. Cambridge: Cambridge University Press.

Ellis, N. C., & Cadierno, T. (2009). Constructing a second language. Annual Review of Cognitive Linguistics, 7 (Special section), 111-290.

Goldberg, A. E. (1995). Constructions. A Construction Grammar Approach to Argument Structure. Chicago: University of Chicago Press.

Goldberg, A. E., Casenhiser, D. M., & Sethuraman, N. (2004). Learning argument structure generalizations. Cognitive Linguistics, 15, 289-316.

Tomasello, M. (2003). Constructing a Language. A Usage-based Theory of Language Acquisition. Cambridge, MA: Harvard University Press.

Roomäe, Kärt

Deciphering Spoken American English: A Construction Grammar Approach to the *how-x* Construction in MICASE

Keywords: spoken English, corpus linguistics, syntax, construction

This work-in-progress report about constructions in spoken English is based on the material from the *Michigan Corpus of Academic Spoken English* (Simpson et al. 2019), abbreviated as MICASE. Not many such corpora exist, and the specific focus of MICASE data on one university's community led me to decide to use this corpus. I am investigating the characteristics of spoken language and the emerging patterns by taking a descriptive approach, with a specific focus on the *how-x* construction.
More specifically, I am looking into dialogic speech events, such as study groups and office hours. All participants are native speakers of American English. For the pilot study, I am using a random sample of 200 data rows downloaded from the corpus. My interest lies in the word *how*, included in all rows, such as (200) *how does it make l- but isn't it, how is it all translated from the same things but it's different?* My preliminary research question is: how do the form-meaning properties lead to the *how-x* construction as attested in MICASE corpus? I will also investigate collocations: looking at the words that follow *how* and in what phrases and clauses they appear, I will analyze in how many rows does *how* form a construction as per the definition, and what instances can only be called sequences.

Considering that it is a spoken corpus, I am expecting to see different sentence structures and deviations from the so-called normative syntax. When analyzing the results I have gotten so far, it has become apparent that my sub-corpus includes, among other features, repetition, ellipsis, non-canonical word order, and discourse particles. These features may affect the usage patterns of the *how-x* construction, its schematic component *x* in particular. Further, I am interested in what utterance types can be found in the dataset, expecting to frequently see questions as *how* is often used as an interrogative adverb which likely will be followed by an auxiliary verb or pronoun. However, declarative and fragmentary utterances may also occur because conversations are collaborative and not all utterances follow canonical structure.

Construction grammar and spoken language have not been connected by researchers working on different languages very frequently, even though some articles do exist (see, e.g., Imo 2005; Michaelis and Feng 2015; Põldvere and Paradis 2019), so my study contributes to theoretical linguistics more broadly.

References:

Imo, W. (2005). A Construction Grammar Approach to the Phrase *I mean* in Spoken English. *Interaction and Linguistic Structures*, 42.

Michaelis, L.A. and Feng, H. (2015). What is this, sarcastic syntax? *Constructions and Frames, 7*(2), 148– 180.

Põldvere, N. and Paradis, C. (2019). 'What and then a little robot brings to you?' The reactive *what-x* construction in spoken dialogue. *English Language and Linguistics,* 1–26. https://doi.org/10.1017/S1360674319000091

Simpson, R.C. et al. (2019). *The Michigan Corpus of Academic Spoken English*. Ann Arbor, MI: The Regents of the University of Michigan. https://quod.lib.umich.edu/cgi/c/corpus/corpus?c=micase;page=simple

Säily, Tanja, Martin Hilpert and Jukka Suomela

New approaches to investigating change in derivational productivity

This paper presents the first results of a project that investigates productivity in historical text corpora by combining constructional and sociolinguistic perspectives. Our focus here is on the nominal suffixes *-ness* and *-ity* in 17th- and 18th-century letters in the *Corpora of Early English Correspondence*. Previous work has indicated that the productivity of the borrowed, more learned and prestigious suffix *-ity* increased during this time and that women were lagging behind in this development, whereas the productivity of the native *-ness* remained stable and nearly free from sociolinguistic variation (Säily 2014; cf. Rodríguez-Puente 2020). This research, however, was based solely on type frequencies, and it ignored several intralinguistic factors of interest. In order to better understand variation and change in the productivity of *-ness* and *-ity*, we analyse the role of four factors, namely etymological type (borrowing/derivative), the word class of the base, branching structure (binary/left/right) and semantics (state/thing/person/collectivity; Romaine 1985). Moreover, we develop new statistical and visual methods that facilitate diachronic comparisons within factors and between competing suffixes (cf. Rodríguez-Puente et al. submitted).

Our pilot results support and refine the earlier finding of a male-led increase in the productivity of *-ity* and provide more information on the timing of the change. Firstly, we use new methods to consider the productivity of *-ity* in relation to *-ness*. We find that *-ity* gains ground on *-ness* with women lagging behind, catching up from the mid-17th century onwards. Secondly, we use these methods to show that the proportion of types that were derived within English increases over time, and women are lagging behind in this development as well (Figure 1). Interestingly, while women exclusively use borrowed types until the end of the 17th century, when they do catch up, they do so quickly, and the overall proportion of derived types only really starts to grow when women join men in using them. Thirdly, the proportion of *-ity* types with adjectival bases increases over time, as does the proportion of 'state' meanings within the derived types. Perhaps counterintuitively from a constructional perspective, these two results also indicate increasing productivity, since the productive use of *-ity* is mostly restricted to adjectival bases, and senses other than 'state' can be argued to be the result of lexicalization (Romaine 1985; Dalton-Puffer 1996:108).

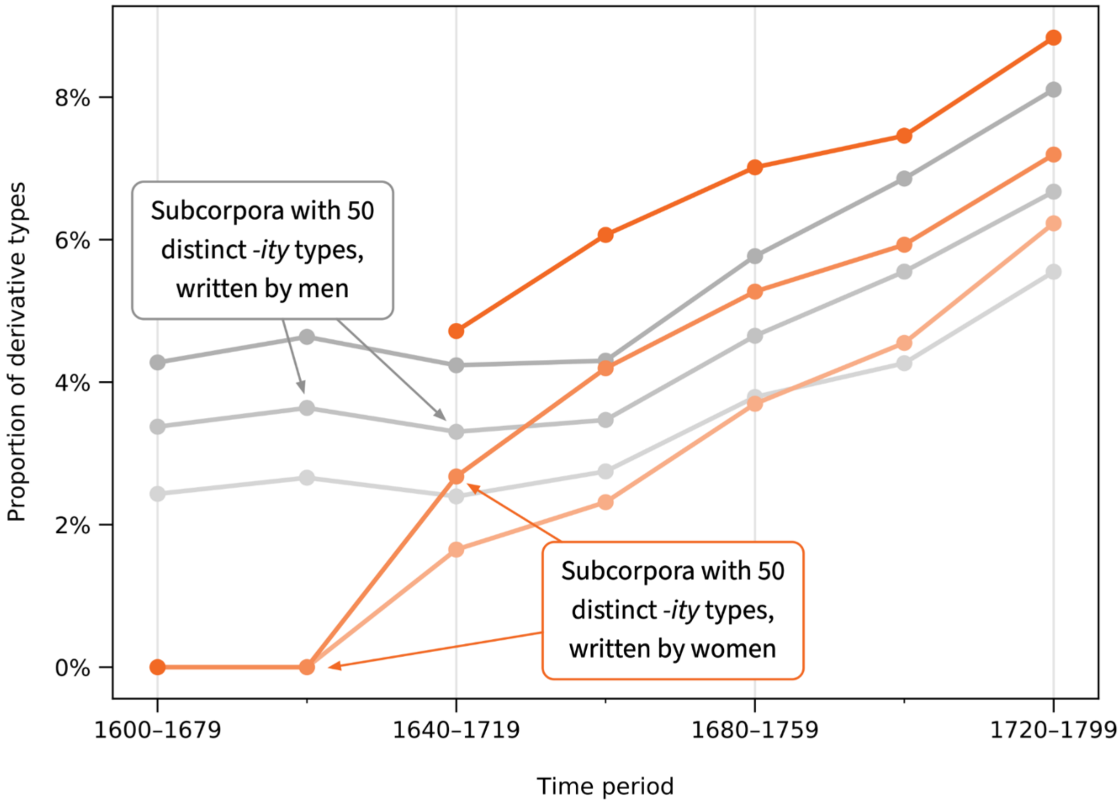

Figure 1. Proportion of types derived within English out of all *-ity* types over time. Sliding window of 80 years, with 20-year increments. Curves: randomly sampled subcorpora with 25/50/75 distinct *-ity* types (for comparability); grey: men; orange: women.

References:

Dalton-Puffer, Christiane. 1996. *The French influence on Middle English morphology: A corpus-based study of derivation*. De Gruyter.

Rodríguez-Puente, Paula. 2020. Register variation in word-formation processes: The development of *-ity* and *-ness* in Early Modern English. *IJES* 20(2):145–167.

Rodríguez-Puente, Paula, Tanja Säily & Jukka Suomela. Submitted. New methods for analysing diachronic suffix competition across registers: How *-ity* gained ground on *-ness* in Early Modern English. *IJCL*.

Romaine, Suzanne. 1985. Variability in word formation patterns and productivity in the history of English. Jacek Fisiak (ed.), *Papers from the 6th International Conference on Historical Linguistics*, 451–465. Benjamins.

Säily, Tanja. 2014. *Sociolinguistic variation in English derivational productivity: Studies and methods in diachronic corpus linguistics*. Société Néophilologique.

Šaldová, Pavlína

Postpositive adjectives and participles: crossing the boundary

In English, the placement of unmodified adjectives (*a black swan* x **a swan black*) after the NP head is highly restricted (the presence of the superlative, temporary attribute). The same holds for past participles, although such restrictions have not been scrutinized in detail. The most productive types in the posthead position are *-ible /-able* adjectives (*available*, *payable*, *applicable,* or *detectable*),, (*un-*)*-ed* adjectives (*interested, appalled, unaddressed*) and past participles of certain transitive verbs (*used, required*, *given* (Furuta 2012)). The paper aims to compare the similarities and differences between the use of adjectives and participles occupying the post-head position. Participles are claimed to occur more freely than adjectives in both positions and "the difference between the two positions [pre- and posthead] in terms of restrictivity is clearest in the case of past participles" (Keizer 2020: 386).

Preferences for and availability in either position in both types of postmodifiers are compared, focussing on identification of factors correlating with and possibly determining the position: type of reference, co-occurrence with determiners and quantifiers, restrictivity, and other discourse-pragmatic factors, such as identifiability of presupposed nature of referent of the NP (James 1979) or the presence/identifiability of a presupposed element (Šaldová 2005).

The *BNC* written component will be used to 1) to identify postpositives ([tag="N.*"] [tag="VVN| AJ0"] [tag="V.*|PU.*"], (previously Blöhdorn (2009) or Furuta (2012)); and 2) to retrieve concordances of the selected items for quantitative and qualitative assessment.

Available as the most frequent light postpositive adjective is unique in the relatively equal proportion between the pre-head, light post-head and heavy-posthead uses, contrasting sharply with other adjectives, where one position dominates (*payable* or *receivable* are very limited as premodifiers), which also holds for certain participles, e.g. for *data collected* vs. *collected data.* On the other hand, both the postpositive participles and adjectives share contexts where quantification plays some role (James 1979), either via explicit presence of quantifiers (*amount* or *number*) or via collocates such as *limit* or *reduce*. Crossing the boundary between postnominal simple adjectives and bare passives as postmodifiers will highlight the shared constraints on the construction.

References:

Blöhdorn, Lars M. 2009. *Postmodifying attributive adjectives in English: an integrated corpus- based approach*. Frankfurt am Main: Peter Lang.
Bolinger, Dwight. 1952. Linear modification. *PMLA* 67. 1117-1144.
Cinque, Guglielmo. 2010. *The Syntax of Adjectives. A Comparative Study*. The MIT Press Cambridge, Massachusetts.

James, Deborah. 1979. Two semantic constraints on the occurrence of adjectives and participles after the noun in English. *Linguistics* 17. 687-705.

Matthews, Peter H. 2014. *The positions of adjectives in English*. Oxford: OUP.
Šaldová, Pavlína. 2005. Presupposition in postmodifying participles: *the assumptions made*. In Jan Čermák, Aleš Klégr, Markéta Malá & Pavlína Šaldová (eds.), *Patterns. A Festschrift for Libuše Dušková*. Praha: Kruh moderních filologů.

Furuta, Y. 2012. Postpositive Past Participles Used on Their Own. *International Journal of Social Science and Humanity*, *2*(6), 514.

Keizer, E. 2020. The problem of non-truth-conditional, lower-level modifiers: a Functional Discourse Grammar solution. *English Language & Linguistics*, *24*(2), 365-392.

Savchenko, Denys

Adjective intensification in the Spoken language of Estonian learners of English

The aim of this work-in-progress is to examine the distribution of intensifiers in the Estonian subcorpus of the Louvain International Database of Spoken English Interlanguage (LINDSEI- EST) in comparison with the English L1 speaker's subcorpus. Quirk et al. (1985: 589-590) define intensifiers as subjuncts that interact with the category of degree and distinguish two main types of intensifiers: amplifiers and downtoners. Ito & Tagliamonte (2003), show that *very*, *really* and *so* are the most frequent intensifiers among British speakers of English. Additionally, the authors point out that the use of intensifiers may function as markers of different generational groups within communities of speakers. Granger (1998) and Lorenz (1999) examined the differences in the use of intensifiers in the written language of both L1 and L2 speakers of English. The results show both underuse and overuse of different types of intensifiers. Considering collocational preferences of intensifiers, it is assumed that intensifiers present certain difficulties for English learners as they may not always be aware of all the factors that determine the choice of adverbs as degree modifiers. Thus, this study is aimed at detailed description of the use of degree adverbs by learners of English in comparison with English L1 speakers.

I will use the data from the Estonian subcorpus of LINDSEI and English L1 equivalent presented in the database. The corpus includes the interviews that are structured around three tasks: three conversation topics, free discussion and a picture description task. All the interviewees in Estonian component were native speakers of Estonian at the third or fourth year of English language and literature programme at the University of Tartu.

The data analysis will require automatic POS-tagging using the Natural Language Toolkit (http://www.nltk.org/) in Python. The focus of this study is intensifiers as modifiers of adjectival heads in both attributive and predicative position. For this, I will extract tokens using tag patterns and chunking rules. The extracted data will be analysed according to frequency and types of items modified by intensifiers. Using collostructional analysis (Stefanowitsch, & Gries 2003), I will identify what adjectives are preferably modified by degree adverbs in Estonian EFL's and English L1 speakers' data. I expect similarities in overall frequencies and differences in types of intensifiers used by EFL speakers.

References:

Ito, Rika and Sali Tagliamonte. 2003. Well weird, right dodgy, very strange, really cool: Layering and recycling in English intensifiers. *Language in Society,* 32: 257–279.

Granger, Sylviane. 1998. Prefabricated patterns in advanced EFL writing: Collocations and formulae. In Anthony Paul Cowie. Phraseology: *Theory, analysis, and applications,* 145–160. Oxford: Clarendon Press.

Lorenz, Gunter. 1999. Adjective Intensification - Learners versus Native Speakers. ACorpus Study of Argumentative Writing. In Jan Aarts and Willem Meijs. *Language and Computers: Studies in Practical Linguistics*, 27. Amsterdam & Atlanta: Rodopi.

Stefanowitsch, Anatol and Stefan Th. Gries. 2003. Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8.2:209-43.

Université catholique de Louvain. (n.d.). Centre for English Corpus Linguistics. LINDSEI. https://uclouvain.be/en/research-institutes/ilc/cecl/lindsei.html

Julia Schlüter

The uptake of corpus evidence on World Englishes by ELT professionals

The ubiquity of variation in language is self-evident to any corpus linguist, and research charting variation in World Englishes has produced important insights over the past two decades. What is more, English serves as an international Lingua Franca between speakers of other languages, raising the stakes in English teaching to unprecedented levels. The fluidity of usage norms thus poses a potential challenge to learners and teachers of English as a Foreign Language. To meet this challenge, the use of corpora has been recommended as "the best tool we can provide future language teachers with" (Granath 2009: 64) as it gives them access to a resource that is more informative than any native speaker.

In an attempt to verify this strong claim, the present study explores how professional language teachers respond to corpus data on examples of questionable usage. The hypothesis – inspired by a variationist perspective – assumes that corpus evidence for the commonness of a variant in one or more varieties of English will increase acceptance of this variant among language professionals.

To test this hypothesis, this paper reports on a meta-study involving language teachers from German secondary schools and universities (both native and non-native speakers of English) in a bipartite rating task. Initially, participants were asked to carry out a routine correction task focussing on dubious prepositional usage. In a second round, the task description was identical, but prior to their ratings participants were exposed to a visual display of corpus data on the item in question.

As test cases, 33 examples of attested divergence between British and American English (e.g. enrolled on/in a course, at/on short notice, different to/from, in respect of/with respect to, at/in a pinch, in/of two minds; Algeo 2006: 159-198) were subjected to quantitative study in 20 varieties, based on the corpus of Global Web-based English (GloWbE, Davies 2013). Example sentences were drawn and adapted from the British Academic Written English Corpus (BAWE) and the Michigan Corpus of Academic Spoken English (MICASE).

The results reveal that the hypothesis cannot be maintained with the given degree of generality. The uptake of descriptive corpus data by language professionals differs vastly, depending on factors such as native/non-native speaker status, age, level of education, level of students taught as well as idiosyncratic attitudes. In the absence of any instructions on or discussion of how the information is to be interpreted and what the implications for the assessment of student writing could be, a considerable number of participants end up being less accepting of variants than before exposure to corpus data.

References:

Algeo, John (2006) British or American English? A Handbook of Word and Grammar Patterns. Cambridge: Cambridge University Press.

Davies, Mark. (2013) Corpus of Global Web-Based English: 1.9 billion words from speakers in 20 countries (GloWbE). Available online at https://www.english-corpora.org/glowbe/.

Granath, Solveig (2009) Who benefits from learning how to use corpora? In: Karin Aijmer (ed.) Corpora and Language Teaching. Amsterdam/New York: Benjamins. 47-65.

Schneider, Gerold

Digital Dickens: an automatic content analysis of Charles Dickens' novels

We follow this year's conference topic of "crossing boundaries through corpora" and use corpora to combine corpus linguistics, stylistics, Digital Humanities, literature and history. As material, we use a corpus of eight influential novels by Charles Dickens and compare them to the voices of contemporary writers from the period. Charles Dickens is famous for his social criticism, for describing and the topic of poverty and for his visions of inclusion of the poor into society. Our first RQ is thus whether and how an automated analysis of his style, topics and content can bring this criticism to the surface. What are, for instance, the associations of poverty in Dickens' novels? We employ methods of overuse, distributional semantics (Sahlgren 2006) and topic modelling (Blei 2012) to investigate the context of descriptions of poverty. These contexts reveal signals of understanding and sympathy, opposed to associations of disgust, disease and crime in his contemporaries, for which we use the CLMET corpus (De Smet 2005).

Dickens is a representative of literary realism. Can this style be traced by automated methods, can we explore the rich imagery that is constructed? This is our second research questions. Particularly topic modeling reveals meticulous descriptions of body language (Mahlberg 2013), cozy dinner scenes, coach travels, detailed descriptions of landscapes, but also his irony. Dickens' style is highly technical, which leads to a dense vocabulary and a high noun/verb ratio (e.g. Pennebaker et al. 2014).

If an automated analysis should be useful for the teaching and scientific exploration of novels, we need to find ways to plot their plots. Our third research question is how successfully this task can be achieved. We test distributional models that convert a text into a conceptual map (McClure 2015). We find the landscapes arising in the process very inspirational, an invitation to wander in interpretations of the novels, but difficult to evaluate.

As no gold standard exists for the data-driven methods that we employ, our approach has to stay partly exploratory and cannot be fully evaluated. This is on the one hand a disadvantage, on the other hand the methods reveal prototypical texts to zoom in, thus supporting the dialectic oscillation between distant reading and close reading (Moretti 2013).

References:

Blei, David. 2012. Probabilistic Topic Models. *Communications of the ACM*, 55, 4, 77--84.

De Smet, Hendrik. 2005. A corpus of Late Modern English. *ICAME Journal* 29: 69-82.

Mahlberg, Michaela. 2013. *Corpus Stylistics and Dickens's Fiction*. Routledge Advances in Corpus Linguistics Series, 14. New York and London: Routledge.

McClure, David. 2015. Textplot Refresh. http://dclure.org/tutorials/textplot-refresh/ Accessed November 30, 2020.

Moretti, Franco. 2013. *Distant Reading*. London: Verso.

Pennebaker James W., Cindy K. Chung, Joey Frazee, Gary M. Lavergne and David I. Beaver. 2014. When Small Words Foretell Academic Success: The Case of College Admissions Essays. *PLoS ONE* 9(12): e115844.

Sahlgren, Magnus. 2006. *The Word-Space Model: Using distributional Analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces.* Ph.D. thesis, Stockholm University.

Sugiura, Masatoshi , Daisuke Abe and Yoshito Nishimura

L1/L2 Differences in Terms of Mean Syntactic Distance (MSD) found in Learner Corpus Data: Some Evidence for the Shallow Structure Hypothesis

The Shallow Structure Hypothesis, claiming that “L2 speakers tend to have problems building or manipulating abstract syntactic representations,” has been one of the thought- provoking hypotheses in SLA (Clahsen & Felser, 2006, 2018). If the real-time syntactic processing is shallow, the depth of the processed products of the sentence structures in
corpus data should be shallow and less complex as well.

The present study proposes mean syntactic distance (MSD) as a measure of syntactic complexity. Similarly to mean hierarchical distance (Jing & Liu, 2015), this measure takes the average distance from the topmost node of a syntactic tree to each
other node. For example, the sentence “John loves Mary.” can be parsed as
[ROOT [S [NP [NNP John]] [VP [VBZ loves] [NP [NNP Mary]]]]]
(1+2+2+3+3+3+4)/7 = 2.57
The average distance from the ROOT node to all other seven non-lexical nodes (S, NP, VP, NNP, VBZ, NP, NNP) is calculated by
To compare MSDs of L1 and L2 syntactic structures, we compiled three comparable
datasets of argumentative essays written on the same topics, 88,000 words each. Two of these were written by Japanese EFL learners (JP-A and JP-B) and the other was written by
native English speakers (NS).

We parsed all the sentences with Stanford Parser. JP-A and JP-B contained 6,615 sentences, including 276,702 nodes, and 6,730 sentences, including 277,183 nodes respectively. NS contained 3,248 sentences, including 243,077 nodes.
With these data, we calculated MSDs: 5.33(JP-A), 5.43(JP-B), and 7.64 (NS).
A series of Kruskal-Wallis tests showed that the MSD of NS was significantly higher than JP-A and JP-B, while JP-A and JP-B did not show a significant difference against each other. The same results were obtained in the analyses for the nodes.

These results suggest that the structures produced by learners are syntactically
shallower than those produced by native speakers.

References:

Clahsen, H., & Felser, C. (2006). Grammatical processing in language learners. Applied Psycholinguistics, 27, 3–42.

Clahsen, H., & Felser, C. (2018). Some notes on the shallow structure hypothesis. Studies in Second Language Acquisition, 40(3), 693-706. doi:10.1017/S0272263117000250

Jing, Y., & Liu, H. (2015). Mean hierarchical Distance augmenting mean dependency distance. Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015), 161-170. Uppsala University, Uppsala, Sweden. https://www.aclweb.org/anthology/W15-2119

Sun, Shuyi and Peter Crosthwaite

A corpus-based cross-disciplinary study of *negation* in Ph.D. theses' limitations sections

There is a consensus among English for Academic Purposes (EAP) scholars that thesis writers' abilities to effectively use interpersonal language, engage with alternative views, and establish solidarity with their disciplinary community is a key feature of successful Ph.D. thesis writing (Hyland, 2006). As a high-stakes, indispensable part-genre in Ph.D. theses across most disciplines, the 'limitations' section contains writers' caveats about their findings, methods or claims as realized through *negation* (e.g. "could not be generalizable"), functioning to convince disciplinary expert examiners to view any shortcomings more favorably (Paltridge & Starfield, 2020). *Negation*, as a *disclaim* marker within Martin and White's (2005) *appraisal* framework, combines with other features to help construct writer-reader relationships at the discourse-semantic stratum (Hood, 2006). However, the discipline-specific role of *negation* in theses' 'limitations' sections remains underexplored.

Accordingly, this study explores *negation* via the *appraisal* framework, addressing subtypes of *negation* (*disalignment*, *cautious detachment*, *unfulfilled expectation*, *validity*) within the 'limitations' sections of Ph.D. theses across disciplines (hard-applied, hard-pure, soft-applied, soft-pure). Our research questions are:
(1) What features of *negation* are employed in the 'limitations' sections of Ph.D. theses?
(2) What is the extent of disciplinary variation in the forms and functions of *negation* within
'limitations' sections?
We collected 120 'limitations' sections of Ph.D. theses to compile four corpora of 30 texts per corpus representing each disciplinary group following Becher's (1989) soft/hard applied/pure typology. UAMCorpusTool (O'Donnell, 2019) was adopted to annotate *negation* subcategories alongside other relevant *appraisal* resources (subtypes of *attitude*, *engagement*, *graduation*). Cross-corpus variation in the target resources were then calculated and visualized using R (R Core Team, 2020).

Our findings showed notable variation in using *negation* across soft-applied and hard-applied corpora, e.g. novice writers in soft-applied disciplines more frequently adopted *negation* alongside *attitude* of *in/security* (anxiety/confidence) via *inscribed* (direct), *negative* modes. Inter- disciplinary variation was also identified across soft-applied and soft-pure corpora with less *negation* found in the latter, though writers in soft-pure disciplines more frequently *upscale* (intensify/enhance) *inscribed attitude* compared with writers in hard disciplines. Analyses of *negation* sub-categories co-occurring with other *appraisal* features (e.g. *entertain* + *negation*:*cautious_detachment*) further indicated cross-disciplinary variation in the choice and sequence of co-articulated items and their rhetorical effects, revealing insights into writers' discipline-specific ways of presenting their work's limitations. We close by explaining how the findings can inform disciplinary thesis writing practice and the incorporation of a corpus-based approach to EAP studies.

References:

Becher, T. (1989). *Academic tribes and territories: Intellectual inquiry and the culture of disciplines*. Open University Press.

Hood, S. (2006). The persuasive power of prosodies: Radiating values in academic writing. *Journal of English for Academic Purposes*, *5*(1), 37–49.

Hyland, K. (2006). *English for academic purposes: An advanced resource book*. Routledge. Martin, J. R., & White, P. (2005). *The language of evaluation: Appraisal in English*. Palgrave
Macmillan.

O'Donnell, M. (2019). UAMCorpusTool [Computer software]. http://www.corpustool.com/index.html

Paltridge, B., & Starfield, S. (2020). *Thesis and dissertation writing in a second language*. Routledge.

R Core Team. (2020). *R* [Computer software]. https://www.r-project.org/

Tagliamonte, Sali and Jeremy Needle

Honest to Pete! Honesty expressions in contemporary North American English

Even though speakers are generally assumed to tell the truth (e.g. Grice 1975), expressions of honesty are widely used in language. In addition to claiming sincerity, these expressions serve a variety of pragmatic functions including softening negative information, asserting independence of opinions, among others (Edwards & Fasulo 2006; Hamilton, Vohs, & McGill 2014). Not surprisingly, the study of honesty expressions has been focused on such pragmatic influences. However, honesty expressions involve a variable array of adverbs, adverbial constructions and collocations which straddle the boundary between morphology and pragmatics, as (1); often alternating from one expression to another, as (1c), showing inherent variation and potentially layering over time.

(1) (a)

2. (b) This is the *god honest truth* ...
3. (c) I can *honestly, truthfully* say that I enjoyed my high-school years.

Some expressions are ubiquitous (*to be honest*); others are non-standard (*honest to Pete),* while shifts in preferences suggest recycling and innovation (*frankly* vs. *seriously).* What is the nature of variation in this area of language and can we establish its linguistic characteristics and possible trajectory of change?

Our data come from a multi-million-word corpus of vernacular North American speech stratified by age, sex and other social factors. The analysis was restricted to unambiguous forms with parallel usage in conversation for a total of 1305 tokens from speakers born between 1884– 2004. Preliminary results reveal the adverb *honestly* dominates (36% of tokens) and among other forms are: *truthfully, God honest, seriously*. Using conditional inference trees as an exploratory analysis tool (see Hothorn, Hornik, & Zeileis 2006) confirms that speaker year of birth is the most important predictor. Moreover, there is a complex interplay of forms at different points in (apparent) time. For speakers born between 1903 and 1958, there is a gendered difference: women use more *honestly*, as well as more *to tell the truth* and *truly* collocations. This gender effect for *honestly* increases for speakers born between 1958 and 1981, where *honestly* is even more dominant for women, but men show a balance between *honestly* and *to be honest* collocations. For speakers born 1981-1994, a sudden spike in the use of *seriously* emerges yet recedes between 1994-2004, giving way to *honestly* and *to be honest*. Mixed effects regression modelling corroborates that year of birth has a significant positive effect on the use of *honestly,* yet social factors are overarchingly stable outside of the limited periods of acceleration of innovating forms.

These findings demonstrate that pragmatic functions are engaged in systemic change and further, that the boundaries of linguistic variables are not restricted to specific categories of grammar (see also Brook 2018). Indeed, pragmatic functions work together with grammatical categories (i.e. adverbs) in ongoing linguistic change. We will extend the analysis to internal factors (e.g. syntactic position) and discuss the methodological and theoretical implications for future analyses of variation in other adverbial forms and expressions.

Oh my god, it's a baby bear. *Honest*!

Gries, Stefan Th.

A new approach to (key) keywords analysis: augmenting frequency-based keywords using dispersion

A widely-used method in corpus-linguistic approaches to discourse analysis, register/genre analysis, and educational/curriculum questions is keywords analysis, a method aiming to identify words that are key to, i.e. characteristic for, certain discourses, text types, or topic domains. Most keywords analyses relied on the same measure as most collocation studies, the log-likelihood ratio G^2 computed from frequencies of occurrence in two corpora under consideration.

To improve on this simple/default mode of analysis and incorporate the distribution of the a words in the target corpus, two main suggestions were made: (i) Baker (2004) proposes to include only those keywords that occur in a sufficiently high number/proportion of texts in the target corpus, and (ii) the method of key key words, words "that show up as key in a large number of texts from the target corpus" (see Egbert & Biber 2019:92). Egbert and Biber then propose an approach that involves computing G^2 s for word types based on the ranges, not the frequencies, of their distribution in the target and reference corpora under consideration.

Their approach is a most welcome addition to keywords analysis, but is too simplistic because it uses (i) only dispersion, not frequency *and* dispersion and (ii) the most simplistic dispersion measure, range, which distinguishes neither sizes of corpus parts nor words' frequencies in corpus parts. Here, I propose to use dispersion in a way that addresses both these problem as well as the problem that G^2 reflects both the frequency and the effect size of the measured keyness. Specifically, I present a new 2-dimensional approach in which keyness is split up into a frequency and a dispersion component, which are both operationalized using the information-theoretic measure of the KL-divergence, a measure quantifying how different one probability distribution is from another: The keyness of a word *w* in terms of

- − frequency is measured by how much *w*'s frequencies in the target and the reference corpus differ from these corpora's sizes;
- − dispersion is measured by how evenly distributed *w* is across the target corpus: all other things being equal, a word *w* more evenly distributed in the target corpus should be more key than a word more clumpily distributed in the target corpus.

I then discuss results of two case studies on the Clinton-Trump corpus and the British National Corpus (academic writing vs. rest). They show that (i) the traditional G^2 approach loses too much information whereas (ii) the current one allows researchers to easily identify which dimension(s) of a word are really responsible for keyness in the G^2 approach, (iii) dispenses with the need for stoplists, and (iv) neatly distinguishes general academic-writing keywords (high- frequency, even-dispersion such as *defined*, *similarly*, *degree*, *factors*, *significance*, *extent*, *analysis*, ...) and domain-specific keywords (high-frequency, uneven-dispersion such as *crohn*, *colorectal*, *χ*, *oesophageal*, *pylori*, *colonic*, *labov*).

Baker, P. 2004. Querying keywords. *JrnlEnglLing* 32(4.) 346-359.

Egbert, J & Biber, D. 2019. Incorporating text dispersion into keyword analyses. *Corpora* 14(1). 77-104.

Schneider, Edgar W.

Lexicosemantic variability in World Englishes: a case study of prospective verbs

Research on World Englishes has produced many descriptions of features of these varieties on the levels of grammar, phonology and loan vocabulary. However, a topic which has practically not been investigated so far is lexicosemantic variability: Is there systematic variation between meanings of words in new varieties? Comparing lexicosemantic variability invites two perspectives: polysemy and semantic fields. Many English words are polysemic, and their individual meanings tend to correlate with characteristic context factors such as complementation patterns, collocating lexemes, or semantic constraints imposed on their role partners. Semantic fields (the focus of investigation in this paper) consist of lexemes which sub-divide a given semantic space amongst themselves, also employing contextual factors like syntactic behavior. Do postcolonial varieties maintain the semantic and contextual properties of their donor variety? Preliminary evidence from a pilot study (Author 2020), briefly reported here, suggests that this is not always the case and that a focusing process may be taking place in some varieties, with rare meanings further reduced in usage (or given up) but core meanings increasingly strengthened.

The present paper investigates lexicosemantic variability in a semantically coherent group of verbs (a word field) in two "metropolitan" (British and American English) and five postcolonial varieties (English in India, Singapore, Hong Kong, the Philippines, and Nigeria). Extrapolating from an earlier corpus-based analysis (Author 1988), "prospective" verbs, those expressing thoughts about possible future events and states, have been chosen, including the lexemes intend, expect, look forward to, plan, contemplate, anticipate, envisage, and envision. All tokens of these verbs have been extracted from the national components of the "International Corpus of English" (ICE) project, for the countries listed above (with the Santa Barbara Corpus substituting the lacking oral component of ICE-USA). Ultimately, 2477 tokens (achieved after some manual pruning) have been coded for four core factors (and other categories not documented here): variety, lemma, meaning (distinguishing six prototypical meanings found in earlier research to subdivide the prospective semantic space), and verb complementation type. Correlations between these individual categories are documented and tested for statistical significance. Effects of and interactions between these nominal variables, paying particular attention to the possible impact of variety, are explored overall, employing three statistical modeling techniques in R: a hierarchical configural frequency analysis (HCFA; Gries (2004), Conditional Inference Trees, and Random Forests. These analyses are supplemented by some qualitative exemplification of potentially innovative (embryonic) construction types.

Results document systematic interrelationships between lexemes, meanings and complementation patterns, and also some significant interactions involving national varieties, which can be viewed as emergent schematic constructions in specific varieties. Overall, the paper provides significant insights into the lexicosemantic variability of meanings and their contextual conditions in World Englishes, thus developing an innovative perspective on linguistic variation and evolution.

References:

Author 1988.

Author 2020.

Gries, Stefan Th. 2004. HCFA 3.2 – A Program for Hierarchical Configurational Frequency Analysis for R for Windows.

Weckermann, Michelle

Prepositions in a cognitive linguistics framework: A cross-linguistic comparison of semantic networks between German and English prepositions

This paper investigates prepositions in a cognitive linguistics framework, including image schemas and semantic networks to represent their polysemy. There is a wide range of research on this subject, including the polysemy of English prepositions (e.g. Hanazaki, 2005 for 'by'; Tyler & Evans, 2003 for 'over'), prepositions in different languages (e.g. Luraghi, 2009 for Italian 'da'; Meex, 2001 and Bellavia, 1996 for German 'über'), as well as cross-linguistic comparative studies (e.g. Taylor, 1988). However, only a small number of papers have constructed semantic networks to illustrate the polysemy of the prepositions, and many have based their analysis on made-up examples (e.g. Tyler & Evans, 2003; Lakoff, 1987).

This paper investigates the polysemy of the two prepositions 'about' and 'after' (among others, as this paper is part of a larger project looking at a range of prepositions) that have not received much attention in previous research. The aim is to construct semantic networks mirroring the polysemy of the two prepositions. Data on the prepositions' different senses was gathered from a selection of corpora, reflecting how the prepositions are employed in natural language. This included a legal corpus (EuroParl), as well as four novels from different genres (thriller, romance/drama, dystopia/fantasy, and philosophical novels) in order to ensure that the data is extracted from a range of topic areas that mirror as many of the different nuances of meaning manifested in the prepositions' senses as possible.

Furthermore, a cross-linguistic dimension was added to the analysis of the two prepositions. Using the same methodological approach, data was gathered for their translation counterparts in German ('um' and 'nach', respectively). The aim of this was to compare and contrast the semantic networks of the prepositions in English and German, observing how the prepositions behave cross-linguistically. Specifically, the goal was to look at whether prepositions are idiosyncratic in their polysemy or whether cross-linguistic similarities in the prepositions' senses can be spotted.

The analysis revealed that, while there are cross-linguistic similarities in the networks of the prepositions (e.g. 'after' and 'nach' both mostly have temporal and spatial senses), there are also interesting differences. 'About', for instance, translates as both 'über' and 'um' in German. Where English thus has one semantic network for 'about' covering both its relational and spatial senses, German has two separate networks for these sense groups. The German network for 'über' furthermore shows overlap with the English network for 'over'.

This paper thus builds on previous research in three main ways: by investigating more and different prepositions (i.e. prepositions that have not been the focus of a lot of research yet), by employing natural and authentic corpus data, and by adding a cross-linguistic dimension to the semantic network-based analysis of the prepositions.

Weihs, Claus and Sarah Buschfeld

RePrInDT: Resampling in unbalanced corpora

Studies in linguistics are often characterized by data sets with unbalanced class variables. In particular, quantitative variationist research investigating the use of non-standard speech forms often faces these problems since such forms occur at much lower token frequencies than standard speech forms (e.g. Buschfeld 2020). This presentation shows ways to statistically meet the problems of unbalanced data sets that are characterized by low token frequencies in the small class. We discuss different resampling methods regarding their capability to generalize from observed tokens in the selected subset of the data to those tokens not selected for modelling and, ideally, to similar tokens that are not even part of the corpus. This is considered to be the generalization capability of a model.

In the present study, we draw on a corpus of L1 child Singaporean and British English. The data from Singapore come from 30 children of different ethnicities, male and female, aged 2;5 (two years; five months) to 12;1. The data from England come from 13 monolingual and 8 bilingual children aged 2;1 to 10;9. All data were elicited by means of video-recorded task- directed dialogue between researcher and child, consisting of a grammar elicitation task, a story retelling task, elicited narratives, and free interaction. The recorded material was orthographically transcribed and manually coded for the realization of subject pronouns, i.e. realized vs. zero.

Since the studied classification problem is heavily unbalanced, the baseline classification rule in the two-class case is “always take the larger class,” which generates an accuracy rate of around 90% in our study. In order to find classification rules that are able to adequately classify the smaller class, Weihs and Buschfeld (2021) employed undersampling of the larger class in their PrInDT approach (Prediction and Interpretation in Decision Trees). The methodological approach developed and pursued in this paper is an extension of PrInDT, in which we also undersample the smaller class by employing different percentages (plarge and psmall) of undersampling for the two classes. In order to adequately represent the accuracies of both classes for model evaluation, we employ the so-called balanced accuracy measure that determines the mean of the accuracies of the two classes. Measured on the overall data set, the best-balanced accuracy is 70.3%, which is reasonably high for such differences in class size.

The results of the study show that the intralinguistic predictor PRONOUN TYPE has the strongest influence on the realization of subjects, followed by MLU (mean length of utterance) and the extralinguistic predictors AGE, LINGUISTIC BACKGROUND, ETHNICITY, and SEX. In general, Singaporeans, and in particular those of Chinese descent, use zero subjects more frequently with pronoun *it* than children growing up in England. Moreover, bi- and multilingualism, male gender, and very young age are predictors for the use of zero subjects. Thus, we have found a model of comparatively high predictive accuracy that clearly depicts how language use is influenced by biological and sociolinguistic factors.

References:

Buschfeld, S. 2020. *Children’s English in Singapore: Acquisition, Properties, and Use*. London: Routledge.

Weihs, C., Buschfeld, S. 2021. Combining Prediction and Interpretation in Decision Trees (PrInDT) - a Linguistic Example. arXiv: http://arxiv.org/abs/2103.02336.

Weilinghoff, Andreas

Pushing transcription work to the next level – Using ASR and LaBB-CAT for linguistic studies

The transcription of sound data is an essential yet time-consuming and labour-intensive part of almost all linguistic research projects. This is especially true for corpus linguistics, as only well transcribed and carefully annotated corpora provide a reliable basis for subsequent analyses. As recent years have seen great advancements in the fields of data mining and speech technology, including techniques such as Automatic Speech Recognition (ASR) (Yu and Deng 2015; Watanabe et al. 2017), a central question is how researchers can incorporate such tools to speed up and enhance transcription work.

My software demonstration addresses this fundamental question. I will show practical and user-friendly strategies of how ASR technology and data mining tools can be effectively used for different study purposes. After a brief theoretical background and a quick overview of different ASR systems, the presentation incorporates a live demonstration of the IBM Watson Speech to Text service applied on a regional variety of English. I will outline strengths and limitations of the service and I will show how ASR text output can be adapted for different transcription software applications, including Praat and ELAN.

Furthermore, I will also demonstrate how broad transcriptions and ASR output can be implemented into LaBB-CAT, a browser-based corpus management and data mining tool developed at New Zealand Institute of Language, Brain and Behaviour (Fromont 2019). LaBB-CAT stands out due to its user-friendly interface and its compatibility with other linguistic software packages. Thus, I will show how broad transcriptions can be automatically annotated, forced aligned, corrected, searched and exported by using different reference dictionaries, software extensions as well as regular expression commands in LaBB-CAT. I will also provide help where technical difficulties may arise.

First investigations and tests have shown that the implementation of ASR technology and subsequent data preparation via LaBB-CAT have proven to be very effective for many linguistic studies. The approach can speed up and enhance transcription work in many research contexts. No coding skills are required for this software demonstration, even though a general understanding of programming languages is advantageous.

References:

Fromont, R. & Hay, J. (2020): *LaBB-CAT*. [Software]. Retrieved from: http://labbcat.sourceforge.net/ [Date of access: 25 November 2020]

Fromont, R. (2019): *Forced Alignment of Different Language Varieties Using LABB-CAT. Proceedings of the 19th International Congress of Phonetic Sciences.*

IBM Watson Speech to Text (2020): [Software]. Retrieved from: https://www.ibm.com/cloud/watson-speech-to-text [Date of access: 25 November 2020].

Watanabe, S., Delcroix, M., Metze, F. & Hershey, J.R. (2017): *New Era for Robust Speech Recognition – Exploiting Deep Learning*. Springer International Publishing.

Yu, D. & Deng, L. (2015): *Automatic Speech Recognition – A Deep Learning Approach.* Springer International Publishing.

All resources for this talk can be found at https://andreas-weilinghoff.com/ASR.html