ICAME43

ANGLIA RUSKIN UNIVERSITY

July 27 – 30, 2022 – Anglia Ruskin University Cambridge

Book of Abstracts

BLOOMSBURY

sense street

amazon

PLENARIES **1**

CORPUS PHONOLOGY: EXPLORING STANDARD SCOTTISH ENGLISH ULRIKE GUT (UNIVERSITY OF MÜNSTER) 2
WHAT CAN WE KNOW ABOUT A WORD FROM THE COMPANY IT KEEPS? A CRITICAL APPRAISAL OF CORPUS-BASED SEMANTICS ALESSANDRO LENCI 3
COMPUTATIONAL MODELLING OF SHORT-TERM LEXICAL SEMANTIC CHANGE IN CONTEMPORARY ENGLISH BARBARA MCGILLIVRAY (KING'S COLLEGE LONDON) 4
'CALM DOWN DEAR!' WOMEN'S LINGUISTIC PARTICIPATION IN UK POLITICAL INSTITUTIONS SYLVIA SHAW (UNIVERSITY OF WESTMINSTER) 6

CONTRASTIVE WORKSHOP PAPERS **7**

HELP CONSTRUCTIONS IN ENGLISH AND NORWEGIAN WITH INFINITIVE COMPLEMENTS THOMAS EGAN (INLAND NORWAY UNIVERSITY OF APPLIED SCIENCES) 8
CONCESSIVE SUBORDINATION IN ENGLISH AND NORWEGIAN HILDE HASSELGÅRD (UNIVERSITY OF OSLO) 10
CROSS-LINGUISTIC DEPENDENCY LENGTH MINIMIZATION IN SCIENTIFIC LANGUAGE: COMPARING ENGLISH AND GERMAN IN THE LATE MODERN PERIOD MARIE PAULINE KRIELKE (SAARLAND UNIVERSITY) 12
ENGLISH NOUN PHRASE COMPLEXITY IN CONTRAST – THE CASE OF HYPHENATED PREMODIFIERS IN NON-FICTION MAGNUS LEVIN & JENNY STRÖM HEROLD (LINNAEUS UNIVERSITY) 15
COMPLEXITY AND NON-FINITE CLAUSES: COMPARING ENGLISH AND GERMAN USAGE HANNA MAHLER (ALBERT-LUDWIGS-UNIVERSITÄT FREIBURG) 17
THE IMPACT OF THE INTENDED READER ON LANGUAGE COMPLEXITY: A CONTRASTIVE VIEW OF SUPPLEMENTIVE PARTICIPIAL CLAUSES IN CHILDREN'S FICTION MARKÉTA MALÁ (CHARLES UNIVERSITY, PRAGUE) 19
STRUCTURAL AND SEMANTIC FEATURES OF ADJECTIVES ACROSS LANGUAGES AND REGISTERS SIGNE OKSEFJELL EBELING (UNIVERSITY OF OSLO) 21
MOOD AND MODALITY: THE SPANISH SUBJUNCTIVE AND ITS ENGLISH COUNTERPART(S) ROSA RABADÁN & NOELIA RAMÓN (UNIVERSITY OF LEÓN) 23
STYLISTIC REPETITION IN NON-FICTION WRITING: CONTRASTIVE AND TRANSLATIONAL PERSPECTIVES JUKKA TYRKKÖ (LINNAEUS UNIVERSITY) 25

PAPERS **27**

A CONTRASTIVE STUDY OF -ISH IN ENGLISH AND SWEDISH KARIN AIJMER (UNIVERSITY OF GOTHENBURG) KARIN.AIJMER@SPRAK.GU.SE 28
THE *HANSARD CORPUS*: SEMANTICS AND SCAFFOLDING MARC ALEXANDER, FRASER DALLACHY, EWAN HANNAFORD (UNIVERSITY OF GLASGOW) 30
KWIC PATTERNS: A NEW NORMAL FOR DISPLAYING, ORDERING, AND INTERPRETING CONCORDANCE LINE RESULTS LAURENCE ANTHONY (WASEDA UNIVERSITY) ANTHONY@WASEDA.JP 32
THE FUTURE OF WORLD ENGLISHES: WILL VERSUS BE GOING TO IN THE INTERNATIONAL CORPUS OF ENGLISH AXEL BOHMANN (UNIVERSITY OF FREIBURG) AXEL@BOHMANN.DE 33
LEXICAL SOPHISTICATION IN SPOKEN ENGLISH: LEX COMPLEXITY TOOL AND THE SPOKEN BNC2014 WORDLIST RAFFAELLA BOTTINI & VACLAV BREZINA (LANCASTER UNIVERSITY) 34
WHILE THE BEACON-FIRE BLAZED ITS BRIGHTEST, THE TWO WOMEN SHRIEKED THEIR LOUDEST: ON THE SUPERLATIVE OBJECT CONSTRUCTION (SOC) TAMARA BOUSO 36

SEMI-STABLE SYSTEMS IN PDE: PARADIGMATIC ENRICHMENT OF CONSTRUCTIONALPARADIGMS LIESELOTTE BREMS (UNIVERSITY OF LIÈGE, RESEARCH FELLOW KU LEUVEN) LBREMS@ULG.AC.BE 38
VISUALIZING ENGLISH LANGUAGE: SYNCHRONIC AND DIACHRONIC TRENDS VACLAV BREZINA & RAFFAELLA BOTTINI (LANCASTER UNIVERSITY) RAFFAELLABOTTINI@GMAIL.COM 40
DEMENTIA METAPHORS IN THE BRITISH PRESS: A CORPUS-BASED STUDY GAVIN BROOKES (LANCASTER UNIVERSITY) 41
...BECAUSE THE LAW COMMANDS IT. A SOCIOLINGUISTIC STUDY OF CAUSAL CONJUNCTIONS IN THE OLD BAILEY CORPUS THOMAS BRUNNER (CATHOLIC UNIVERSITY OF EICHSTAETT) THOMAS.BRUNNER@KU.DE 42
HOW 'REAL' IS THE QUANTITATIVE TURN? INVESTIGATING STATISTICS AS THE 'NEW NORMAL' INCORPUS LINGUISTICS SARAH BUSCHFELD (TU DORTMUND UNIVERSITY), SARAH.BUSCHFELD@TU-DORTMUND.DE SVEN LEUCKERT (TECHNISCHE UNIVERSITÄT DRESDEN) ANDREAS WEILINGHOFF (TU DORTMUND UNIVERSITY) CLAUS WEIHS (TU DORTMUND UNIVERSITY), CLAUS.WEIHS@TU-DORTMUND.DE 44
DIACHRONIC ANALYSIS OF GRAMMATICAL FORMS AND FUNCTIONS IN A CORPUS OF 16TH- TO 19TH-CENTURY ENGLISH GRAMMAR BOOKS BEATRIX BUSSE, NINA DUMRUKCIC, SOPHIE DU BOIS, & INGO KLEIBER 45
THE OXFORD COMMA IN THE HISTORY OF ENGLISH JAVIER CALLE-MARTÍN (UNIVERSITY OF MÁLAGA) & MIRIAM CRIADO-PEÑA (UNIVERSITY OF GRANADA) JCALLE@UMA.ES, MCRIADOP@UGR.ES 47
RESEARCH TRENDS IN CORPUS LINGUISTICS: A BIBLIOMETRIC ANALYSIS OF TWO DECADES OF SCOPUS- INDEXED CORPUS LINGUISTICS RESEARCH IN ARTS AND HUMANITIES INTRODUCTION PETER CROSTHWAITE (UNIVERSITY OF QUEENSLAND), P.CROS@UQ.EDU.AU SULISTYA NINGRUM (INDONESIA/STATE POLYTECHNIC OF SRIWIJAYA), ARUM.EFFENDI@GMAIL.COM MARTIN SCHWEINBERGER 48
CORPUS LINGUISTICS MEETS THE LAW: CAN AN AMERICAN PRESIDENT ONLY BE IMPEACHED FOR CRIMINAL CONDUCT? CLARK D. CUNNINGHAM (GEORGIA STATE UNIVERSITY COLLEGE OF LAW), CDCUNNINGHAM@GSU.EDU UTE RÖMER (GEORGIA STATE UNIVERSITY), UROEMER@GSU.EDU 50
COMBINING CORPUS AND QUALITATIVE METHODS TO IMPROVE THE REPRESENTATION OF SPOKEN LANGUAGE IN ELT MATERIALS NIALL CURRY (COVENTRY UNIVERSITY) 51
LINGUISTIC FACTORS IN SUCCESSFUL PERSUASION ONLINE DARIA DAYTER (TAMPERE UNIVERSITY) 53
THE CRISIS OF NORMALITY. CONCEPTUAL METAPHORICAL PATTERNS IN THE DISCOURSE OF CRISIS: NEW OLD MAPPINGS DARIO DEL FANTE (UNIVERSITY OF PADOVA) 55
REPORTING CLAUSES IN BRITISH GENERAL VS. CRIME FICTION JARLE EBELING & SIGNE OKSEFJELL EBELING JARLE.EBELING@GMAIL.COM; S.O.EBELING@ILOS.UIO.NO 57
DETERMINING LETTER-SPECIFIC SPEECH ACTS IN 18TH CENTURY VARIETIES OF ENGLISH CHRISTINE ELSWEILER & PATRICIA RONAN (TU DORTMUND UNIVERSITY) PATRICIA.RONAN@TU-DORTMUND.DE 59
CAPTURING AND ANALYSING MULTIMODALITY IN A CORPUS OF ROYAL CORRESPONDENCE: A CASE STUDY OF THE LETTERS OF JAMES V AND HENRY VIII MEL EVANS (UNIVERSITY OF LEEDS) & HELEN NEWSOM (ASTON UNIVERSITY) M.EVANS5@LEEDS.AC.UK; H.NEWSOME@ASTON.AC.UK . 61
A CORPUS-BASED ANALYSIS OF IRISH ENGLISH SPEAKERS' VIRTUAL INTERCULTURAL COMMUNICATIONS IN THE TECHNOLOGY SECTOR GAIL FLANAGAN (UNIVERSITY OF LIMERICK) GAIL.FLANAGAN@UL.IE 63

A QUASI-LONGITUDINAL ANALYSIS OF THE L2 ACQUISITION OF TENSE AND ASPECT ROBERT FUCHS & VALENTIN WERNER ROBERT.FUCHS.DD@GMAIL.COM 65
THE DATIVE ALTERNATION IN PRE- AND POST-HANDOVER HONG KONG: TOWARDS ENDONORMATIVE STABILIZATION OR RESTRICTION? NINA FUNKE (JUSTUS LIEBIG UNIVERSITY GIESSEN) NINA.FUNKE@ANGLISTIK.UNI-GIESSEN.DE 67
TRIANGULATING METHODS IN CORPUS LINGUISTICS: FROM FREQUENCY TO MOVE AND DIALOGIC ANALYSIS MATTEO FUOLI (UNIVERSITY OF BIRMINGHAM) & MONIKA BEDNAREK M.FUOLI@BHAM.AC.UK 69
ESTABLISHING A 'NEW NORMAL': DETECTING FLUCTUATING TRENDS IN WORD FREQUENCY OVER TIME ANDREW KEHOE, ANDREW.KEHOE@BCU.AC.UK MATT GEE, MATT.GEE@BCU.AC.UK ANTOINETTE RENOUF, ANTOINETTE.RENOUF@BCU.AC.UK (BIRMINGHAM CITY UNIVERSITY) 70
PHRASAL VERBS AS MULTIWORD UNITS: A COMPARISON OF EFL AND ESL GAËTANELLE GILQUIN (UNIVERSITY OF LOUVAIN) GAETANELLE.GILQUIN@UCLOUVAIN.BE 72
EXPLAINING REGIONAL PATTERNS IN MORPHOSYNTACTIC DIALECT FEATURES: THE CASE OF BE *SAT/STOOD* IN ENGLAND AND BEYOND JASON GRAFMILLER (UNIVERSITY OF BIRMINGHAM), J.GRAFMILLER@BHAM.AC.UK JACK GRIEVE 74
MOST DISPERSION MEASURES DO NOT MEASURE DISPERSION STEFAN TH. GRIES (UCSB & JLU GIESSEN) STGRIES@GMAIL.COM 79
MANY ASSOCIATION MEASURES DO NOT MEASURE ASSOCIATION (BUT FREQUENCY), AND WHAT TO DO ABOUT THAT STEFAN TH. GRIES (UCSB & JLU GIESSEN), STGRIES@GMAIL.COM MAGALI PAQUOT 81
ARE SNOWCLONES THE NEW NORMAL? USING CORPORA TO STUDY EXTRAVAGANT FORMULAIC PATTERNS STEFAN HARTMANN (UNIVERSITY OF DÜSSELDORF) & TOBIAS UNGERER HARTMAST@HHU.DE 82
EXPLORING EMERGING PATTERNS OF SELF-IDENTIFICATION IN THE *LGBTQ+ REDDIT CORPUS* TURO HILTUNEN, TURO.HILTUNEN@HELSINKI.FI LAURA HEKANAHO MINNA PALANDER-COLLIN HELMIINA HOTTI 84
RHYTHM IN WORLD ENGLISHES – EVIDENCE FROM A QUANTITATIVE ANALYSIS OF CO-OCCURRENCE PATTERNS IN CORPORA OF L1 AND L2 VARIETIES OF ENGLISH SEBASTIAN HOFFMANN (TRIER UNIVERSITY), HOFFMANN@UNI-TRIER.DE SABINE ARNDT-LAPPE (TRIER UNIVERSITY), ARNDTLAPPE@UNI-TRIER.DE PETER UHRIG (FAU ERLANGEN-NÜRNBERG), PETER.UHRIG@FAU.DE 86
PROMOTION AND PRESERVATION OF PUBLIC HEALTH: TRENDS IN HEALTH SCIENCE AND COMMUNICATION IN THE ROYAL SOCIETY CORPUS KATHERINE IRELAND (UNIVERSITY OF GEORGIA) KATHERINE.IRELAND@UGA.EDU 88
GRAMMATICAL NATIVIZATION IN SPOKEN SOUTH ASIAN ENGLISHES: THE CASE OF THE EXISTENTIAL-THERE CONSTRUCTION KATHRIN KIRCILI (UNIVERSITY OF MARBURG), KATHRIN.KIRCILI@UNI-MARBURG.DE JULIA DEGENHARDT, JULIA.DEGENHARDT@ADMIN.UNI-GIESSEN.DE TOBIAS BERNAISCH SANDRA GOETZ 90
A CORPUS-BASED ACOUSTIC ANALYSIS OF VOWEL PRODUCTION BY L1-JAPANESE LEARNERS AND NATIVE SPEAKERS OF ENGLISH YUKI KOMIYA (THE UNIVERSITY OF QUEENSLAND), Y.KOMIYA@UQCONNECT.EDU.AU MARTIN SCHWEINBERGER 92
A COMPLEX PUZZLE: COMPARING THEORY-BASED MODELS OF GRAMMATICAL COMPLEXITY IN SPOKEN VERSUS WRITTEN REGISTERS TOVE LARSSON (NAU - NORTHERN ARIZONA UNIVERSITY),TOVE.LARSSON@NAU.EDU DOUGLAS BIBER GREGORY R. HANCOCK 94
ADVERB PLACEMENT IN L2 SPOKEN PRODUCTION: THE EFFECT OF LINGUISTIC AND EXTRALINGUISTIC FACTORS TOVE LARSSON (NAU- NORTHERN ARIZONA

University),tove.larsson@nau.edu Marcus Callies, Tülay Dixon Hilde Hasselgård (University of Oslo), hilde.hasselgard@ilos.uio.no Nicole Hober, Natalia Judith Laso, Isabel Verdaguer, Sanne van Vuuren, Magali Paquot 96
'Guv, I'm a copper, not a social worker!': Using corpus-assisted discourse studies to analyse how caring professionals are portrayed on anglophone TV Maria Leedham (The Open University) Maria.leedham@open.ac.uk 98
"I shall be glad if you will note…" – Studying early 20th century business correspondence from Hong Kong to assess variety-specific genre developments Lisa Lehnen (University of Würzburg), lisa.lehnen@uni-wuerzburg.de Ninja Schulz (University of Würzburg), ninja.schulz@uni-wuerzburg.de Carolin Biewer 100
Semantic reanalysis and idiomatization: multi-word verbs in the Late Modern English period Ljubica Leone (Lancaster University) l.leone1@lancaster.ac.uk.....102
Grammaticalization of Aspect in German and its diachronic parallels in English Zlata Liwschin (Leibniz University of Hannover) zlata.liwschin@germanistik.uni-hannover.de.....104
"Like English is used everywhere" – The functions and use of discourse marker like in UAE English Eliane Lorenz (Justus Liebig University Giessen, Norwegian University of Science and Technology) eliane.lorenz@anglistik.uni-giessen.de.....106
Challenges in deriving a new COLT from the Spoken BNC2014: the case of teenage swearing Robbie Love (Aston University) & Anna-Brita Stenström r.love@aston.ac.uk 108
This essay argues that: Connecting metadiscursive nouns and rhetorical moves in research article abstracts Liu Luda (Jilin University) 110
The avoidance of repetition in translation: A multifactorial study of repeated reporting verbs in the Italian translation of the Harry Potter series Lorenzo Mastropierro.....111
Is death the great equalizer? A study of news accounts of women and men as murder victims Monika Mondor (Gothenburg University) & Joe Trotta (Gothenburg University) monika.mondor@sprak.gu.se; joe.trotta@sprak.gu.se.....113
Do formulaic sequences mask proficiency? Considering evidence from a large learner corpus Akira Murakami, a.murakami@bham.ac.uk Ute Römer (Georgia State University), uroemer@gsu.edu Marije MichelDora Alexopoulou 115
***Not that you asked, but here it is*: a formal and functional taxonomy of *not-that* clauses in Present-day American English** Ozan Mustafa (University of Graz) ozan.mustafa@uni-graz.at 117
Concerns about cancer immunotherapy in online forum posts: A corpus-based discourse analysis Hoa Ninh.....119
Investigating Sentiments on Covid-19 in Tweets Niklas Nitsch (TU Dortmund University) & Patricia Ronan (TU Dortmund University) niklas.nitsch@tu-dortmund.de, patricia.ronan@tu-dortmund.de.....122
A Closer Look at GET in African Postcolonial Englishes Temitayo Olatoye (University of Eastern Finland).....124
Spelling variation in inner-circle Englishes Marta Pacheco-Franco (Universidad de Malaga) martapacheco@uma.es 125
Direct object definiteness and verb meaning: A corpus-based investigation Florent Perek (University of Birmingham) & Lotte Sommerer (University of Freiburg) florent.perek@gmail.com; lotte.sommerer@anglistik.uni-freiburg.de 127

F0 RANGE IN L2 DISCOURSE: A CORPUS-BASED CONTRASTIVE INTERLANGUAGE ANALYSIS KARIN PUGA (JUSTUS LIEBIG UNIVERSITY GIESSEN) KARIN.PUGA@ANGLISTIK.UNI-GIESSEN.DE 129
A GENDER-BASED ANALYSIS OF PRAGMATIC MARKERS IN SRI LANKAN ENGLISH MAHISHI RANAWEERA 131
GRINDING TO A HALT? THE SPREAD OF THE PROGRESSIVE IN RECENT SPOKEN BRITISH ENGLISH PAULA RAUTIONAHO (UNIVERSITY OF EASTERN FINLAND) PAULA.RAUTIONAHO@UEF.FI 133
ANOTHER TURN OF THE SCREW ON THE DEVELOPMENT OF *-ITY* AND *-NESS* ACROSS REGISTERS: EARLY AND LATE MODERN ENGLISH PERIODS IN FOCUS PAULA RODRÍGUEZ-PUENTE(UNIVERSITY OF OVIEDO) RODRIGUEZPPAULA@UNIOVI.ES 135
WHAT'S NORMAL IN CONVERSATIONS? NON-CANONICAL INTERROGATIVES IN LINDSEI-EST CORPUS KÄRT ROOMÄE (UNIVERSITY OF BIRMINGHAM) KXR177@STUDENT.BHAM.AC.UK 137
WORD ORDER IN ADDITIONAL-LANGUAGE ENGLISH SPOKEN BY MULTILINGUALS SYLVI RØRVIK (INLAND NORWAY UNIVERSITY OF APPLIED SCIENCES) SYLVI.RORVIK@INN.NO 139
READ THIS POLICY: A CORPUS-BASED ANALYSIS OF TERMS OF USE CONTRACTS TIM SAMPLES, TSAMPLES@UGA.EDU CAROLINE KRACZON KATHERINE IRELAND (UNIVERSITY OF GEORGIA), KATHERINE.IRELAND@UGA.EDU 141
***ONCE UPON A TIME, THERE WAS A FAIRY TALE … AND IT LIVED HAPPILY EVER AFTER*: A CONTRASTIVE CORPUS-BASED STUDY OF GERMAN AND ENGLISH FAIRY TALE OPENINGS AND CLOSINGS** CHRISTINA SANCHEZ-STOCKHAMMER (TU CHEMNITZ), ASYA YURCHENKO (TU CHEMNITZ) CHRISTINA.SANCHEZ@PHIL.TU-CHEMNITZ.DE; ASYA.YURCHENKO@PHIL.TU-CHEMNITZ.DE 144
TRUMP'S POPULIST RHETORIC: A CORPUS-BASED ANALYSIS OF 'THE PEOPLE', AND 'THE ELITE' JULIA SCHILLING (UNIVERSITY OF HAMBURG) JULIA.SCHILLING@UNI-HAMBURG.DE 146
PANDEM-ONIUM: IDENTIFYING KEYWORDS AND PHRASES IN BRITISH COVID-19 TWITTER AND NEWSPAPER DISCOURSE JULIA SCHILLING (UNIVERSITY OF HAMBURG) & ROBERT FUCHS JULIA.SCHILLING@UNI-HAMBURG.DE; ROBERT.FUCHS.DD@GMAIL.COM 148
FLUENCY IN ASIAN ENGLISHES: A MULTIVARIATE CORPUS-BASED ANALYSIS OF INDIAN AND SRI LANKAN ENGLISH KAROLA SCHMIDT (JLU GIESSEN) KAROLA.SCHMIDT-1@ANGLISTIK.UNI-GIESSEN.DE 150
NETWORK VISUALISATIONS OF LINGUISTIC RELATIONSHIPS IN LARGE DATASETS: A CASE STUDY EXPLORING THE CONTEXT OF *NORMAL* IN BRITISH ENGLISH HANNA SCHMÜCK (LANCASTER UNIVERSITY) H.SCHMUECK@LANCASTER.AC.UK 152
SEPARATING THE WHEAT FROM THE CHAFF: HOW TO DETECT IDIOMS VS. COMPOSITIONAL COLLOCATIONS WITH COLLOCATION MEASURES AND DISTRIBUTIONAL SEMANTICS GEROLD SCHNEIDER (UNIVERSITY OF ZURICH) GSCHNEID@IFI.UZH.CH 155
A NEGATOR WALKS INTO A MODAL CLAUSE A DIACHRONIC CORPUS-BASED STUDY OF A COMPLEX RELATIONSHIP ULRIKE SCHNEIDER (UNIVERSITY OF MAINZ) ULRIKE.SCHNEIDER@UNI-MAINZ.DE 157
COMPILING A DIACHRONIC CORPUS TO TRACE VARIETY-SPECIFIC GENRE CONVENTIONS ACROSSTIME: CHALLENGES AND SOLUTIONS FOR AUTOMATISING TEXT RECOGNITION OF BUSINESSCORRESPONDENCE FROM HONG KONG NINJA SCHULZ (UNIVERSITY OF WÜRZBURG), NINJA.SCHULZ@UNI-WUERZBURG.DE LISA LEHNEN (UNIVERSITY OF WÜRZBURG), LISA.LEHNEN@UNI-WUERZBURG.DE CHRISTIAN REUL & CAROLIN BIEWER.....160
***ALL*-CLEFT CONSTRUCTIONS IN THE LONDON–LUND CORPUS 2 (LLC–2) OF SPOKEN BRITISH ENGLISH** ELENI SEITANIDI (LUND UNIVERSITY), ELENI.SEITANIDI@ENGLUND.LU.SE NELE PÕLDVERE & CARITA PARADIS 161

STATING THE OBVIOUS: ASSUMED EVIDENTIAL PARENTHETICALS WITH *VERBA DICENDI* IN CONTEMPORARY ENGLISHES MARIO SERRANO-LOSADA (COMPLUTENSE UNIVERSITY OF MADRID), MARIO.SERRANO@UCM.ES ZELTIA BLANCO-SUÁREZ (UNIVERSITY OF SANTIAGO DE COMPOSTELA), ZELTIA.BLANCO@USC.ES 163
COMPILING A CORPUS OF SOUTH ASIAN ONLINE ENGLISHES: SOME REFLECTIONS AND A PILOT STUDY MUHAMMAD SHAKIR (WWU MUENSTER) & DAGMAR DEUBER MUHAMMADSHAKIRAZIZ@OUTLOOK.COM 165
ATTITUDES TO IMMIGRATION IN AUSTRALIAN HANSARD: 1970-2020 ADAM SMITH (MACQUARIE UNIVERSITY) & MINNA KORHONEN (MACQUARIE UNIVERSITY) ADAM.SMITH@MQ.EDU.AU 167
KEYWORD ANALYSIS: PROGRESS THROUGH REGRESSION LUKAS SÖNNING (UNIVERSITY OF BAMBERG) LUKAS.SOENNING@UNI-BAMBERG.DE 169
SEEING THE WOOD FOR THE TREES: RECURSIVE PARTITIONING WITH MARGINAL EFFECTS PLOTS LUKAS SÖNNING (UNIVERSITY OF BAMBERG) & JASON GRAFMILLER (UNIVERSITY OF BIRMINGHAM) LUKAS.SOENNING@UNI-BAMBERG.DE; J.GRAFMILLER@BHAM.AC.UK 171
MODELING 21ST-CENTURY BANGLADESHI ENGLISH THROUGH CORPUS DATA CRISTINA SUÁREZ-GÓMEZ CRISTINA.SUAREZ@UIB.ES 173
EPISTEMIC VERB EXPRESSIONS IN NATIVE AND NON-NATIVE WRITING: TASK EFFECTS (WORK IN PROGRESS) DAISUKE SUZUKI (UCL) DAISUKE.SUZUKI.19@UCL.AC.UK 175
AMERICANIZATION IN INDIVIDUAL FINNISH LINGUA FRANCA ENGLISH USER'S NETWORKS. WORK IN PROGRESS IRENE TAIPALE (UNIVERSITY OF EASTERN FINLAND) IRENE.TAIPALE@UEF.FI 177
PURPOSE SUBORDINATORS ON THE MOVE: *SO, SO THAT, JUST SO* NETWORK ELNORA TEN WOLDE (UNIVERSITY OF GRAZ), ELNORA.TEN-WOLDE@UNI-GRAZ.AT GUNTHER KALTENBÖCK (UNIVERSITY OF GRAZ), GUNTHER.KALTENBOECK@UNI-GRAZ.AT 179
SHOULD'VE DID A CORPUS STUDY, COULD'VE WROTE A PAPER JOE TROTTA (GOTHENBURG UNIVERSITY) & MONIKA MONDOR (GOTHENBURG UNIVERSITY) JOE.TROTTA@SPRAK.GU.SE; MONIKA.MONDOR@SPRAK.GU.SE 181
BIG THYME FOR BIG DATA METHODS - APPROACHING THE QUESTION OF HOMOPHONE DURATIONS WITH A LARGE AUTOMATICALLY ANNOTATED DATASET PETER UHRIG (FRIEDRICH-ALEXANDER-UNIVERSITÄT ERLANGEN-NÜRNBERG) PETER.UHRIG@FAU.DE 182
THE ADAPTATION OF A CORPUS: REFORMATTING CANBEC FOR SOCIOLINGUISTIC ANALYSIS ISOLDE VAN DORST (VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS) ISOLDE.VAN.DORST@WU.AC.AT 184
ANALYSING CATEGORY CHANGE WITH ENRICHED DATA: A GRAMMATICAL AND SOCIOLINGUISTIC STUDY OF -*ED* PARTICIPLES, 1810–2009 TURO VARTIAINEN & TANJA SÄILY (UNIVERSITY OF HELSINKI) TANJA.SAILY@HELSINKI.FI 186
CHUNKING AT THE INDIVIDUAL LEVEL SVETLANA VETCHINNIKOVA (UNIVERSITY OF HELSINKI) SVETLANA.VETCHINNIKOVA@HELSINKI.FI 188
ON THE IMPORTANCE OF THE COMMON GROUND IN SCIENTIFIC DISCOURSE: EXTENDERS ANDFOCUS OPERATORS IN LATE MODERN ENGLISH IRIA BELLO VIRUEGA & ESTEFANÍA SÁNCHEZ BARREIRO (UNIVERSIDADE DA CORUÑA) FANNILLASB@GMAIL.COM 190
ACCURATE CONFIDENCE INTERVALS ON BINOMIAL PROPORTIONS, FUNCTIONS OF PROPORTIONS AND RELATED SCORES SEAN WALLIS (SURVEY OF ENGLISH USAGE - UCL) S.WALLIS@UCL.AC.UK 193
"SITTING *ON* A CHAIR WRITING A PAPER *ON* PREPOSITIONS": A COGNITIVE SEMANTIC STUDY OF THE POLYSEMY OF THE PREPOSITION *ON* MICHELLE WECKERMANN (UNIVERSITY OF AUGSBURG) MICHELLE.WECKERMANN@PHILHIST.UNI-AUGSBURG.DE 194

PRAGMATIC VARIATION IN WORLD ENGLISHES: A CORPUS-PRAGMATIC ANALYSIS OF QUESTION TAG USE IN NIGERIAN, PHILIPPINE, AND TRINIDADIAN ENGLISH MICHAEL WESTPHAL (UNIVERSITY OF MÜNSTER) MICHAEL.WESTPHAL@WWU.DE 196
DISCOURSES OF 21ST CENTURY IDENTITY DOCUMENTS IN THE UK: A CONTRIBUTION TO DIACHRONIC CORPUS STUDIES VIOLA WIEGAND V.WIEGAND@BHAM.AC.UK 198
AGREEMENT WITH COLLECTIVE NOUNS IN AFRICAN AND CARIBBEAN ENGLISHES GUYANNE WILSON (TU DORTMUND) GUYANNEWILSON@GMAIL.COM 200
CREOLE AND POWER: A CRITICAL DISCOURSE ANALYSIS OF LEGAL CROSS-EXAMINATIONS IN ICE TRINIDAD AND TOBAGO AND ICE JAMAICA GUYANNE WILSON (TU DORTMUND) & MICHAEL WESTPHAL (UNIVERSITY OF MÜNSTER) GUYANNEWILSON@GMAIL.COM; MICHAEL.WESTPHAL@WWU.DE 202
USE OF EVALUATIVE *THAT* IN RESEARCH ARTICLES: VARIATIONS ACROSS PARADIGMS MEI YANG (UNIVERSITY OF HELSINKI) YSX@HPU.EDU.CN 204
WHICH FACTORS ARE AT PLAY IN ENGLISH ARGUMENT STRUCTURE VARIATION? NPS VS PPS THROUGHOUT TIME EVA ZEHENTNER EVA.ZEHENTNER@ES.UZH.CH 206

POSTERS 208

CONNECTIONS BETWEEN GENRES AND ERROR PATTERNS IN A SWEDISH UPPER-SECONDARY EAL LEARNER CORPUS DANIEL IHRMARK (LINNAEUS UNIVERSITY) DANIEL.O.SUNDBERG@LNU.SE 209
THE DESIGN OF A CORPUS OF LATE MODERN ENGLISH TEXTS ON PHYSICS LUIS PUENTE-CASTELO (UNIVERSIDADE DA CORUÑA), PCASTELO.LUIS@GMAIL.COM LEIDA MARIA MONACO (UNIVERSITY OF OVIEDO), MARIAMONACO86@GMAIL.COM ISABEL MOSKOWICH (UNIVERSIDADE DA CORUÑA), IMOSKOWICH@UDC.ES BEGOÑA CRESPO (UNIVERSIDADE DA CORUÑA),BEGONA.CRESPO.GARCIA@UDC.ES 211
THE IMPACT OF *STAR WARS* ON THE ENGLISH LANGUAGE: A STUDY OF *STAR WARS*-DERIVED WORDS AND CONSTRUCTIONS IN PRESENT-DAY ENGLISH CORPORA CHRISTINA SANCHEZ-STOCKHAMMER (TU CHEMNITZ) CHRISTINA.SANCHEZ@PHIL.TU-CHEMNITZ.DE 213

SOFTWARE DEMOS 215

SOFTWARE DEMONSTRATION: MEANING-BASED QUERYING OF HISTORICAL CORPORA WITH MACBERTH LAUREN FONTEYN & ENRIQUE MANJAVACASL.FONTEYN@HUM.LEIDENUNIV.NL; E.M.A.MANJAVACAS.AREVALO@HUM.LEIDENUNIV.NL 216

Plenaries

Corpus Phonology: Exploring Standard Scottish English

Ulrike Gut (University of Münster)

This talk introduces one of the latest approaches within corpus linguistics: corpus phonology, i.e. the use of phonological corpora for the study of phonetic and phonological properties of languages and accents (e.g. Durand, Gut & Kristoffersen 2014). In the first part, both a brief overview of this new line of research and a description of the components of a phonological corpus (raw data, annotations) will be given. In the second part, corpus phonology will be illustrated with the example of three corpus-based studies investigating Standard Scottish English phonology. Using the phonemically annotated ICE Scotland (Schützler, Gut & Fuchs 2017), rhotics and rhoticity, the realisation of the NURSE vowel as well as the /w/-/ʍ/ contrast in Scottish Standard English will be explored. Results show that this variety of English is variably rhotic, that some rhotics are realised as a tap or trill, and that only few speakers show a full NURSE merger, while few of them maintain the /w/-/ʍ/ contrast. In addition, it was found that the realisation of these phonological features is determined more by language-internal than by social factors and that high variability across speakers exists.

References

Durand, Jacques, Gut, Ulrike & Kristoffersen, Gjert (Eds.) (2014). Oxford Handbook of Corpus Phonology. Oxford: Oxford University Press.

Schützler, Ole, Gut, Ulrike & Fuchs, Robert (2017). New perspectives on Scottish Standard English. Introducing the Scottish component of the International Corpus of English. In Hancil, S. & Beal, J. (Eds.), Perspectives on Northern Englishes (pp. 273-301). Berlin: Mouton de Gruyter.

What can we know about a word from the company it keeps? A critical appraisal of corpus-based semantics

Alessandro Lenci

At least since Firth (1957), corpus data have become a key source of knowledge about word meaning. The idea that we can know a word by analyzing its co-occurrence patterns in language usage has become the central tenet of distributional semantics (Lenci 2018). Much has changed from the early days of corpus-based semantics and highly sophisticated models and methods are now available to acquire information about word meaning, alongside other linguistic knowledge. Actually, corpora have become the only source of semantic information used by Natural Language Processing and Artificial Intelligence systems. In this talk, I will make a general critical appraisal of the achievements and limits of corpus-based semantic methods. What can we extract about word meanings from corpora? Is the new generation of methods, like the most recent neural language models, the real breakthrough that is often heralded? Why do certain aspects of meaning seem to be constantly elusive? Is it a problem of the models or of the data we use? Or does it depend on the way we treat corpus data as a source of meaning? What challenges are in front of us in this research area?

Computational modelling of short-term lexical semantic change in contemporary English

Barbara McGillivray (King's College London)

Lexical semantic change, i.e. the phenomenon in which the semantics of lexical items changes over time, has been the object of qualitative research for over a century. Anthropological studies in linguistics (Boas 1911; Sapir 1912; 1928) and in conceptual history (Williams 1976; Richter 1995) have recognised the importance of this research to reach a better understanding of the dynamics of cultural, social and political systems. Philological methods (e.g., Kenny 1995, Wierzbicka 1997) and theoretical linguistics research (Geeraerts 2010; Koch 2016; Grondelaers et al. 2007) have also engaged with the analysis of language-internal aspects of this phenomenon.

Today, access to very large born-digital collections from the web allows us to study short-term changes in contemporary language data, with the potential to shed new light into our understanding of cultural and societal changes. For example, the verb "follow" acquired the sense of staying informed about someone's postings in 2007, after the launch of the social media platform Twitter. In recent years computational research aimed at detecting lexical semantic change phenomena from large corpora spanning long time intervals has achieved encouraging results (cf. e.g., Schlechtweg et al. 2020), but so far little work has been done on detecting short-term lexical semantic shifts.

In this talk I will present my research on developing computational models for semantic change drawing on state-of-the-art methods and will share my experience of working in a range of interdisciplinary projects dealing with social media and contemporary digital sources, including web archives, Twitter and emoji. This will offer us an opportunity to reflect on the types of lexical semantic changes that can be detected using these methods, and on how they reflect temporary or more lasting changes in contemporary English.

References

Boas, F. (1911). "Introduction", in Boas, F. (ed.), Handbook of American Indian Languages, Washington D.C.: Bureau of American Ethnology, Bulletin 40 (I), pp. 59-73.

Geeraerts, D. (2010). Theories of lexical semantics. Oxford: Oxford University Press

Grondelaers, Stefan, Speelman, Dirk and Geeraerts, Dirk (2007). Lexical variation and change. In The Oxford handbook of cognitive linguistics.

Kenny, Neil. (1995). Interpreting Concepts after the Linguistic Turn: The Example of curiosité in Le Bonheur des sages / Le Malheur des curieux by Du Souhait (1600)', in Interpréter le seizième siècle, ed. by John O'Brien (Michigan Romance Studies, XV, 1996), 241–70.

Koch, P. (2016). Meaning change and semantic shifts. In Päivi Juvonen and Maria Koptjevskaja Tamm (eds.), The Lexical Typology of Semantic Shifts, pages 21–66. De Gruyter Mouton, Berlin/Boston.

Richter, M. (1995). The History of Political and Social Concepts: A Critical Introduction. New York and Oxford: Oxford University Press.

Sapir, E. (1912). "Language and Environment", American Anthropologist 14, pp. 226-242.

Sapir, E. (1928). Proceedings, First Colloquium on Personality Investigation; Held under the Auspices of the American Psychiatric Association, Committee on Relations with the Social Sciences, New York: Lord Baltimore Press, pp. 77-80.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. In Proceedings of the Fourteenth Workshop on Semantic Evaluation, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

‘Calm down Dear!’ Women’s linguistic participation in UK Political Institutions

Sylvia Shaw (University of Westminster)

The underrepresentation of women in politics is a persistent and seemingly intractable problem for British politics. In this paper I discuss some of the findings from my 2020 book, Women, Language and Politics in which I attempt to use theoretical and methodological sociolinguistic approaches to discover some of the obstacles and barriers that women in politics face. Researchers of political institutions are greatly helped by the wealth of data available to them in the analysis of political discourse from video recordings and written records to in-situ observation and large searchable datasets. Here I discuss some of the advantages of combining approaches to identify patterns of participation across different institutions. In doing so, I discuss analyses of parliamentary data from The Scottish Parliament; the Northern Ireland Assembly; the Welsh Assembly and the House of Commons. I also consider some of the pitfalls with conducting gender and language research, including oversimplifying the notion of a ‘women’s’ or ‘men’s style of speech, and failing to recognise the ideological underpinnings of beliefs about gendered behaviour.

Contrastive Workshop Papers

HELP constructions in English and Norwegian with infinitive complements

Thomas Egan (Inland Norway University of Applied Sciences)

Christian Mair has written of the English verb *help* that it is "a corpus linguist's delight" because its distribution in texts is so clearly influenced by stylistic, contextual, semantic and structural constraints, few of which are categorical in the sense that one variant is excluded in a specific environment' (Mair 1995: 261). The same point might well be made with respect to its Norwegian cognate *hjelpe*. The two verbs *help* and *hjelpe* occur in a variety of parallel constructions, often with a very high degree of lexical mutual correspondence (MC: see Altenberg 1999). For instance, in the English–Norwegian Parallel Corpus, which provides the data for the present paper, the construction in which the verb is followed by just one complement, coding a HELPEE, displays an MC of 95% across a total of 77 examples.

This presentation deals with a subset of HELP constructions in the two languages, those containing some form of infinitive complement. In English these are of four types, depending on whether or not the HELPEE is coded explicitly and whether the infinitive is preceded by *to*. In Norwegian there are seven types, differing considerably in complexity, since the infinitive, which is always preceded by the infinitive marker *å* (to), may also be preceded by the HELPEE and by *til* (for), *med* (with), or by both of these in the absence of an explicit HELPEE.

> English constructions: HELPER + [*help*] + (HELPEE) + (*to*) + infinitive…..
> Norwegian constructions: HELPER + [*hjelpe*] + (HELPEE) + (*til*) + (*med*) + *å* + infinitive…..

Much has been written about the English constructions, and possible factors governing the choice between them. One of the most common, though rather controversial, theories states that the more hands-on the role played by the HELPER, the greater the likelihood that the bare infinitive form of the complement is chosen (Wood 1956: 107, Duffley 1992: 25-29, Mair 1995: 262-263, Huddleston & Pullum 2002: 1,244, Dixon 2005: 268, McEnery & Xiao 2005: 169-176, Egan 2008: 207-210). Consider in this respect the following two examples.

(1) "Would you like me to **help** you find a nice one with lots of pictures in it?" (RD1)
— Skal jeg **hjelpe** deg **å** finne noen riktig søte med massevis av bilder i? (RD1T)
— *Shall I help you to find some really sweet with lots of pictures in?*

(2) "You're the exception that **helped** me to make a new start." (ABR1)
"Du er det unntaket som **hjalp** meg **til å** gjøre en ny begynnelse." (ABR1T)
"*You are the exception that helped me for to make a new start.*"

In (1) the HELPER is actively involved in the action of finding, and the bare infinitive construction represents the default choice. In (2), on the other hand, the HELPER is more of an enabler, and here the *to*-infinitive is said to be the more likely choice. In (1) the less complex English construction is translated by the least complex Norwegian one. In (2) the more complex English construction is translated by a complex Norwegian one. In this WiP

I examine the 165 utterances containing infinitival HELP complements in the ENPC with the following goals:

- To shed light on the features (structural/semantic/stylistic) that influence the choice of construction type in the source texts in both languages.
- To chart the correspondences between the construction types in the source and target texts in the two languages, with a particular eye on the degree of complexity of the infinitive complement form.

Preliminary results show that while in the English source texts the constructions with and without an explicit HELPEE are almost equally frequent, the construction with the HELPEE is six times more frequent in the Norwegian source texts than its (structurally) less complex HELPEE-less counterpart. In the Norwegian target texts, the 'HELPEE + *med å* (with to)' construction is most often used to translate examples with hands-on HELPERs, but the 'HELPEE + *til å* (for to)' construction is favoured when the help given is more indirect.

References

Altenberg, Bengt. 1999. Adverbial connectors in English and Swedish: Semantic and lexical correspondences. In Hilde Hasselgård and Signe Oksefjell (eds), *Out of corpora: Studies in honour of Stig Johansson*. Amsterdam: Rodopi, 249–268.

Dixon, Robert M.W.. 1991. *A semantic approach to English grammar. Second edition.* Oxford: Oxford University Press

Duffley, Patrick J.. 1992. *The English infinitive*. London: Longman

Egan, Thomas. 2008. *Non-finite complementation: a usage-based study of infinitive and -ing clauses in English*. Amsterdam: Rodopi.

Huddleston, Rodney and Geoffrey K Pullum. 2002. *The Cambridge Grammar of the English Language.* Cambridge: Cambridge University Press.

Mair, Christian. 1995. Changing patterns of complementation, and concomitant grammaticalisation, of the verb help in present-day English, in Bas Aarts and Charles F. Meyer (eds), *The verb in contemporary English: theory and description*. Cambridge: Cambridge University Press. 258-72.

McEnery, Anthony and Zhonghua Xiao. 2005. *HELP* or *HELP to*: what do corpora have to say? *English Studies*, 86:2:161-187

Wood, Fredrick T.. 1956. Gerund versus infinitive. *English Language Teaching,* 11:11-16.

Concessive subordination in English and Norwegian

Hilde Hasselgård (University of Oslo)

Concessive markers occur when a proposition holds true in spite of expectations or inferences to the contrary (Huddleston & Pullum 2002). Concessive meaning is thereby related to both contrast and condition (Couper-Kuhlen and Kortmann 2000). Cross-linguistic studies of concession in English and Norwegian/Swedish have focused on connectors (Altenberg 2002, Fretheim 2002, Johansson & Fretheim 2002; see also Dupont 2021). The present study therefore concentrates on concessive relations signalled by other types of concessive markers, especially subordinators, exemplified in (1).

(1) *Though* he had little money, he preferred to keep his independence... (AH1)
Selv om han hadde dårlig med penger, foretrakk han å bevare uavhengigheten… (AH1T)

The starting point was *though*, which can function as an adverb, a subordinator and a contrastive particle (*OED*). In the fiction part of the English-Norwegian Parallel Corpus (ENPC), its main Norwegian translation correspondences are *selv om, enda,* and *skjønt* (in that order of frequency), which in turn correspond to *although, even though, even if,* and *though*. Other correspondences include coordinators (chiefly *but/men*) and connectors such as *likevel* ('still', 'after all')*, imidlertid* ('however')*, yet* and *though*. The concessive markers *though, even though, although, even if; selv om, enda* and *skjønt* were selected for further study, addressing the following research questions:

- How do the English and Norwegian concessive markers compare with regard to syntactic functions, both intra- and cross-linguistically?
- To what extent are the concessive markers translated congruently between the languages, and what are the most common noncongruent correspondences?
- What are the positions of the concessive clauses in both languages?

A preliminary analysis indicates that *though*, the most frequent concessive marker in English, typically functions as a subordinator, as in (1), introducing finite, nonfinite and verbless clauses. It also functions as a phrase coordinator, similar to *but*, as in (2), and as an adverbial connector. *Although, even though* and *even if* are always subordinators, as is the most frequent Norwegian marker, *selv om*. However, *enda* can function as a connector, and *skjønt* sometimes has coordinator function (visible from the word order pattern of the clause), as in (3).

(2) But this satisfaction had given way to a worthier *though equally selfish* regret. (MD1)
Men denne tilfredsheten var avløst av en verdigere beklagelse, *skjønt* den var like egoistisk. (MD1T) [Lit: "…to a worthier regret, *though* it was equally selfish."]
(3) *Skjønt* naturligvis hadde hun rett. (EHA1) [Lit: "Though naturally had she right"]
But of course she was right. (EHA1T)

Concessive subordination involves both syntactic and cognitive complexity (Couper-Kuhlen and Kortmann 2000) and may thus represent a challenge for translators. However, the translation paradigms of the markers show a fairly high degree of congruence, although non-congruent correspondences may change subordination to

coordination and phrase-level coordination to clausal subordination, as in (2). The bilingual concordances reveal a complex network of realizations of concessive relations (see also Salkie & Reed 1999; Taboada & Gómez-González 2012).
The position of concessive clauses (Hasselgård 2010) varies language-internally across the subordinators and cross-linguistically. Initial position is more common in Norwegian, especially with *enda*, and end position is more common in English, especially with *though*. Moreover, clauses with *enda, skjønt* and *though* sometimes appear as independent sentences (3), emphasizing the multifunctionality of concessive markers.

References

Altenberg, Bengt. 2002. Concessive connectors in English and Swedish. In *Information Structure in a Cross-Linguistic Perspective,* ed. by Hasselgård, H., Johansson, S., Behrens, B., & Fabricius-Hansen, C., 21–43. Amsterdam: Rodopi. https://doi.org/10.1163/9789004334250_003

Couper-Kuhlen, Elizabeth & Kortmann, Bernd (eds). 2000. *Cause - Condition - Concession - Contrast: Cognitive and Discourse Perspective*. Berlin, New York: De Gruyter Mouton.

Dupont, Maïté. 2021. *Conjunctive Markers of Contrast in English and French. From syntax to lexis and discourse*. Amsterdam: Benjamins.

Fretheim, Thorstein. 2002. Interpreting concessive adverbial markers in English and Norwegian discourse. In *Information Structure in a Cross-Linguistic Perspective,* ed. by Hasselgård, H., Johansson, S., Behrens, B., & Fabricius-Hansen, C., 1–19. Amsterdam: Rodopi.

Hasselgård, Hilde. 2010. *Adjunct Adverbials in English*. Cambridge University Press.

Huddleston, Rodney & Pullum, Geoffrey K. 2002. *The Cambridge Grammar of the English Language*. Cambridge University Press.

Johansson, Stig & Fretheim, Thorstein. 2002. The semantics and pragmatics of the Norwegian concessive marker *likevel*: Evidence from the English-Norwegian Parallel corpus. In L.E. Breivik & A. Hasselgren (eds), *From the COLT's Mouth … and Others'. Language corpora studies in honour of Anna-Brita Stenström*, pp. 81-101. Amsterdam: Rodopi.

Oxford English Dictionaries Online (OED). https://www.oed.com/ (accessed March 2022)

Salkie, Raphael & Oates, Sarah Louise. 1999. Contrast and concession in French and English. *Languages in Contrast*, 2:1, 27-56.

Taboada, Maite & Gómez-González, María de los Ángeles. 2012. Discourse markers and coherence relations: Comparison across markers, languages and modalities. *Linguistics and the Human Sciences*, Vol 6, 17-41.

Cross-linguistic Dependency Length Minimization in Scientific Language: Comparing English and German in the Late Modern Period

Marie Pauline Krielke (Saarland University)

We use Universal Dependencies (UD) for the study of cross-linguistic diachronic change in syntactic complexity in the Late Modern period (ca. 1650 – 1900). We specifically ask whether scientific English and German minimize the length of syntactic dependency relations and if so, how this minimization is implemented. The Late Modern period, marked by the scientific revolution and a turn towards experimental science, is known as the beginning of modern science involving the emergence of a linguistically distinctive register associated with it. While the formation of a new register entails external pressures, e.g., through constantly new emerging vocabulary, factors counterbalancing this pressure to maintain communicative efficiency can be found on the grammatical level (Hawkins, 2004). Previous work has shown that scientific English gradually evolves towards increased lexical density while showing lower grammatical complexity (Halliday, 1988; Biber, 2006; Biber and Gray, 2016). German follows this trend only towards the end of the 19thc. (Möslein, 1974; Beneš, 1981; Admoni, 1990; Habermann, 2011).

In the present study, we use Dependency Length (DL), the linear distance between syntactic heads and their dependents (Hudson, 1995) as a well-studied measure of syntactic complexity correlating with cognitive load as incurred by working memory (Gibson, 1998; Gibson, 2000). To reduce cognitive load, DL is minimized where possible as stated in the Dependency Length Minimization Hypothesis (Gibson, 2000, DLM). This assumption, however, relies on constituency ordering preferences such as short-before-long orderings or heavy-NP shifts (Futrell et al., 2020; Wasow, 2002; Bresnan et al., 2007; Shih et al., 2015). While in synchronic studies languages have shown to optimize word order for DLM (Futrell et al., 2015; Liu et al., 2017), German shows much weaker effects than English, due to word order flexibility (Gildea and Temperley, 2010). In English diachronic studies, DLM was also found for Old and Middle English (Tily, 2010) as a result of optimized word order over time. While studies focusing on the Late Modern English period, too, attest DLM (Lei and Wen, 2020; Juzek et al., 2020), DLM here derives from a trend on the lexico-grammatical with long dependency constructions (clausal embeddings, see (1a) and (2a)) towards short dependency constructions (i.e., noun phrases, see (1b) and (2b)). We are not aware of comparable studies for German.

1) a. *An Observation of* ***a Boy that was Hydropical****.* (Edwardy Tyson, 1683)
   b. *An observation of* ***a hydropical boy****.*

2) a. *Nimm* ***Meisterwurz / die dürr ist*** */ zerschneide sie klein.* (Martin Zeiller, 1659)
   b. *Nimm* ***dürre Meisterwurz****, zerschneide sie kein.*

We look at the Late Modern Period, when major morpho-syntactic changes in English and German had already been completed. To find whether the scientific registers are subject to DLM, we not only look at DL development itself but specifically analyze the frequencies of particular short vs. long dependency relations. Our first hypothesis is that DLM is a cross-lingual optimization process in English and German scientific language in the Late Modern Period. Secondly, we assume that DLM is achieved by an increase in short intra-phrasal dependencies and a decrease in long cross-clausal dependencies. We analyze two large-scale, comparable, UD-annotated corpora (Krielke et al., 2022): the

RSC_UD_parsed_1.0 (parsed Royal Society Corpus, Fischer et al. 2020) and the *DTAW_UD_parsed_1.0* (scientific portion of Deutsches Textarchiv, Geyken et al., 2018). Our macro-analytic results show that both in English and German scientific discourse DL is minimized over time (as observed by 50 years periods). Analyzing different sentence lengths (SL), we find the most significant changes on SL30 and a continuous trend towards lower DL for English, while German DL only decreases significantly in the first and last time periods, which is in line with previous work on scientific English and German. Our micro-analyses show that in both languages, short dependency relations (highly frequent nominal dependents, e.g., determiners and adjectives) become more frequent over time, while most long dependencies (cross-clausal dependencies, e.g., relative clauses) become less favored contributing to overall DLM. Furthermore, we find that cross-clausal dependencies become longer rather than shorter over time. However, the overall impact on DL is negligible since such relations at the same time are either low-frequency throughout or decrease in frequency.

References

Wladimir G. Admoni. 1990. Historische Syntax des Deutschen. Niemeyer.

Eduard Beneš. 1981. Die formale Struktur der wissenschaftlichen Fachsprachen aus syntaktischer Hinsicht. In Theo Bungarten, editor, *Wissenschaftssprache*, pages 185–212. Fink, München.

Douglas Biber and Bethany Gray. 2016. Grammatical Complexity in Academic English: Linguistic Change in Scientific Writing. *Studies in English Language*. Cambridge University Press, Cambridge, UK.

Douglas Biber, 2006. Multi-dimensional patterns of variation among university registers, volume 23 of *Studies in Corpus Linguistics*, chapter 7, pages 177–212. John Benjamins Publishing, Amsterdam/Philadelphia.

Joan Bresnan, Anna Cueni, Tatiana Nikitina, and R. Harald Baayen. 2007. Predicting the Dative Alternation. page 33.

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.

Richard Futrell, Roger P. Levy, and Edward Gibson. 2020. Dependency locality as an explanatory principle for word order. *Language*, 96(2):371–412.

Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Edward Gibson. 2000. Dependency locality theory: A distance-based theory of linguistic complexity. In Y. Miyashita A. Marantz and W. O'Neil, editors, *Image, language, brain: Papers from the first mind articulation project symposium*. Cambridge, MA: MIT Press.

Fischer, Stefan, Menzel, Katrin, Knappen, Jörg, and Teich, Elke. 2020. The Royal Society Corpus 6.0 providing 300+ years of scientific writing for humanistic study. In *Proceedings of the Conference on Language Resources and Evaluation* (LREC). ELRA.

Geyken, Alexander, Boenig, Matthias, Haaf, Susanne, Jurish, Bryan, Thomas, Christian, and Wiegand, Frank. 2018. Das Deutsche Textarchiv als Forschungsplattform für historische Daten in CLARIN. In Henning Lobin, et al., editors, *Digitale Infrastrukturen für die germanistische Forschung*, pages 219–248. De Gruyter

Daniel Gildea and David Temperley. 2010. Do Grammars Minimize Dependency Length? *Cognitive Science*, 34(2):286–310. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1551-6709.2009.01073.x.

Mechthild Habermann. 2011. *Deutsche Fachtexte der Neuzeit. Naturkundlich-medizinische Wissensvermittlung im Spannungsfeld von Latein und Volkssprache*. De Gruyter, Berlin/ Boston.

M.A.K. Halliday. 1988. On the language of physical science. In Mohsen Ghadessy, editor, *Registers of Written English: Situational Factors and Linguistic Features*, pages 162–177. Pinter, London.

J.A. Hawkins. 2004. *Efficiency and Complexity in Grammars*. Oxford linguistics. OUP Oxford.

Richard Hudson. 1995. *Measuring syntactic difficulty*. Manuscript, University College, London.

Tom S. Juzek, Marie-Pauline Krielke, and Elke Teich. 2020. Exploring diachronic syntactic shifts with dependency length: the case of scientific English. In *Proceedings of the Fourth Workshop on Universal Dependencies* (UDW 2020), pages 109–119, Barcelona, Spain (Online), December. Association for Computational Linguistics.

Marie-Pauline Krielke, Luigi Talamo, Mahmoud Fawzi, and Jörg Knappen. 2022. Tracing Syntactic Change in the Scientific Genre: Two Universal Dependency-parsed Diachronic Corpora of Scientific English and German. In *Proceedings of the 13th International Conference on Language Resources and Evaluation* (LREC), Marseille.

Lei Lei and Ju Wen. 2020. Is dependency distance experiencing a process of minimization? A diachronic study based on the state of the union addresses. *Lingua*, 239:102762.

Haitao Liu, Chunshan Xu, and Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of life reviews*, 21:171–193.

Kurt Möslein. 1974. Einige Entwicklungstendenzen in der Syntax der wissenschaftlich-technischen Literatur seit dem Ende des 18. Jahrhunderts. *Zur Geschichte der deutschen Sprache und Literatur*, 94:156–198.

Stephanie Shih, Jason Grafmiller, Richard Futrell, and Joan Bresnan. 2015. Rhythm's role in genitive construction choice in spoken English. *Rhythm in cognition and grammar*, pages 207–234. Publisher: De Gruyter Berlin.

Harry Tily. 2010. *The role of processing complexity in word order variation and change*. Ph.D. thesis, Stanford University.

Thomas Wasow. 2002. Postverbal behavior. Number no. 145 in CSLI lecture notes. CSLI, Stanford, Calif.

English noun phrase complexity in contrast – the case of hyphenated premodifiers in non-fiction

Magnus Levin & Jenny Ström Herold
(Linnaeus University)

This paper continues our previous investigations into complex NP premodification in English, German and Swedish. In 2017, we studied English 'hyphenated premodifiers' such as *this slow-walking* (*figure*) and *fifteen-year-old* (*schoolgirls*) where two or more elements are linked by hyphens to form complex premodifiers (Levin & Ström Herold 2017). That study was based on fiction texts from the *English-Swedish Parallel Corpus* (ESPC) where stylistic effect and author creativity (*T-shirted, cereal-slurping, cartoon-watching Saturday-morning* (*viewers*)) largely guided the formation of these multi-word units (see also Meibauer 2007). In translation, these expressive features also determined many of the equivalents chosen by translators, who largely transferred the source-text structures into the target languages.

However, complex premodification is not only connected to author expressiveness, but perhaps even more so with syntactic condensation (e.g., Biber et al. 1999: 588; Trips 2012: 335). English is undergoing long-term changes in its noun-phrase structures in that increasingly complex material is placed before the noun head (e.g., Biber, Grieve & Iberri-Shea 2009; Smitterberg 2021:187). This "spectacular increase" (Leech et al. 2009: 206) is reflected in premodifying nouns and noun-participle compounds such as *health-related problems* (Biber & Gray 2016: 187–190). Similar changes towards increased premodification have not been ascertained in German and Swedish, but German seems to rely more on complex premodifiers than Swedish (Magnusson 1995), which likely affects the options available to translators.

Our working hypotheses are thus i) that non-fiction genres produce less creative but more numerous hyphenated premodifiers than fiction, ii) that partly different elements are combined into premodifiers in non-fiction compared to fiction, iii) that non-fiction translators use rather different translation strategies, because of the greater focus on content rather than literary style, and iv) that German correspondences to hyphenated premodifiers more often consist of premodifiers than those in Swedish. Our data, about 6,500 English occurrences with German and Swedish equivalents, was retrieved from the *Linnaeus University English-German-Swedish* corpus (LEGS) (Ström Herold & Levin 2021). The corpus consists of recently published non-fiction texts such as popular science, history and self-help books. All texts have been translated into two languages.

Preliminary results suggest that non-fiction texts in LEGS contain more hyphenated premodifiers than fiction in ESPC – compared to ESPC, the frequencies are almost twice as high. Translations into English from German and Swedish originals produce fewer hyphenated premodifiers than English originals, indicating that translated text is less condensed than originals. There are also indications of premodifiers being of a more "stereotypical" nature in non-fiction, partly due to their term-like status (*the short-haired bumblebee*; *armour-piercing shells*). The more technical content of non-fiction is also reflected in a greater use of prefixed premodifiers expressing condensed (semi-)technical content (*quasi-enforced anonymity*; *anti-pollution measures*). Correspondences of English non-fiction premodifiers are postmodifiers (*low-dose lithium* > *litium i låga doser*) slightly more often than in fiction, a tendency which is even stronger in Swedish than in German.

Our study will shed new light on NP modification and condensation in three languages, as well as the strategies available to translators.

References

Biber, Douglas & Bethany Gray. 2016. *Grammatical complexity in academic English. Linguistic change in writing*. Cambridge: Cambridge University Press.

Biber, Douglas, Jack Grieve & Gina Iberri-Shea. 2009. Noun phrase modification. In Günter Rohdenburg & Julia Schlüter (eds.), *One language, two grammars? Differences between British and American English*, 182–93. Cambridge: Cambridge University Press.

Leech, Geoffrey, Marianne Hundt, Christian Mair & Nicholas Smith. 2009. *Change in contemporary English. A grammatical study*. Cambridge: Cambridge University Press.

Levin, Magnus & Jenny Ström Herold. 2017. Premodification in translation: English hyphenated premodifiers in fiction and their translations into German and Swedish. In Egan, Thomas & Hildegunn Dirdal (eds.), *Cross-linguistic correspondences. From lexis to genre*. Amsterdam: Benjamins. 149–175.

Magnusson, Gunnar. 1995. Deutsch–Schwedisch kontrastiv: Stolpersteine bei avancierter Übersetzung. *Moderna språk* 89(2), 164–179.

Meibauer, Jörg. 2007. How marginal are phrasal compounds? Generalized insertion, expressivity, and I/Q-interaction. *Morphology* 17(2): 233–259.

Smitterberg, Erik. 2021. *Syntactic change in Late Modern English*. Cambridge: Cambridge University Press.

Ström Herold, Jenny & Magnus Levin. 2021. The colon in English, German and Swedish: A contrastive corpus-based study. In Rössler, Paul, Peter Besl & Anna Saller (eds.), *Comparative Punctuation – Vergleichende Interpunktion*. Berlin: De Gruyter. 237–261.

Trips, Carola. 2012. Empirical and theoretical aspects of phrasal compounds: Against the "syntax explains it all" attitude. *On-line Proceedings of the Eighth Mediterranean Morphology Meeting*(MMM8). [http://www.unistuttgart.de/linguistik/sfb32/fies/mmm8_proceedingsteil1.pdf]

Complexity and non-finite clauses: comparing English and German usage

Hanna Mahler (Albert-Ludwigs-Universität Freiburg)

Non-finite clauses, being a more compressed and less explicit subtype of dependent clauses (Biber & Gray 2016: 207, 219), play an important role when discussing (clausal) grammatical complexity. When contrasting non-finite clauses in English and German we do, however, also find considerable differences in the complexity of the system of non-finite verb phrases.

While the structural inventories of both languages are well described (e.g. König & Gast 2012), and claims as to the greater reliance on non-finite clauses in English abound (e.g. Rohdenburg 1990: 151), a comprehensive, systematic comparison of actual usage patterns is to date still missing. Existing research either focuses on translation strategies for (specific types of) non-finite clauses (e.g. Fabricius-Hansen 1998, Ström Herold & Levin 2018) or is based on limited data (e.g. Fischer 2013). This project therefore strives to:

- provide empirical support for existing hypotheses about English-German contrasts,
- analyse to what degree which constructions contribute to the perceived differences in the frequency of non-finite clauses,
- investigate the role of register and medium for the contrasts in the use of non-finite clauses,
- explore the cross-linguistic relationship between the frequency of non-finite clauses and overall information density and complexity of a text.

To provide a concrete example: both English and German allow postmodification of noun phrases through infinitives, present participles, and past participles (for example: The house *built* in the year 1900. The house *standing* far from the village. The house *to be sold*. Das Haus, in 1900 *erbaut*. Die Studie, *aufbauend* auf vorherige Experimente. Der Versuch, pünktlich *zu sein*.); the structural possibilities are therefore quite comparable. Nevertheless, previous research indicates that non-finite postmodifying clauses are considerably more frequent in English than in German (e.g. König & Gast 2012: 243). On the other hand, English has many more possibilities of including explicit subjects in present and past participial clauses (Biber et al. 1999: 125), but this option appears to not be utilised too often (Biber et al. 1999: 198). This short illustration already reveals the limitations of a purely system-based comparison and the necessity of a usage-based comparison considering medium- und register-differences.

The study at hand uses GECCo, a comparable corpus of spoken and written texts from English and German (Kunz et al. 2021). Through a combination of automatic and manual processing (which is currently being implemented), all verb phrases in the corpus are identified and annotated for their finiteness, their verb form, and their grammatical function. The frequency of finite and non-finite verb phrases (in various functions) in English and German can then be examined with the help of mixed-effects regression modelling.

The study will therefore also be able to assess the contribution of non-finite verb phrases to the assumed "verbality" of English compared to the German "nominality" (Kortmann & Meyer 1992: 163).
Taking the contrastive perspective therefore helps to reveal the language-specific ways in which non-finite constructions are employed as speakers balance information density, linguistic complexity, and economy.

References

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman Grammar of Spoken and Written English.* Harlow: Pearson Education Limited.

Biber, Douglas & Bethany Gray. 2016. *Grammatical Complexity in Academic English: Linguistic Change in Writing* (Studies in English Language). Cambridge: Cambridge University Press.

Fabricius-Hansen, Cathrine. 1998. Informational Density and Translation, with Special Reference to German - Norwegian - English. In Stig Johansson & Signe O. Ebeling (eds.), *Corpora and Cross-Linguistic Research: Theory, Method, and Case Studies* (Language and computers 24), 197-234. Amsterdam, Atlanta: Rodopi.

Fischer, Klaus. 2013. *Satzstrukturen im Deutschen und Englischen: Typologie und Textrealisierung* (Konvergenz und Divergenz 1). Berlin: Akademie Verlag.

König, Ekkehard & Volker Gast. 2012. *Understanding English-German Contrasts,* 3rd edn. (Grundlagen der Anglistik und Amerikanistik 29). Berlin: Erich Schmidt Verlag.

Kortmann, Bernd & Paul Meyer. 1992. Is English Grammar More Explicit than German Grammar, after all? In Christian Mair & Manfred Markus (eds.), *New Departures in Contrastive Linguistics: Proceedings of the Conference Held at the Leopold-Franzens-Universität Innsbruck, Austria, 10-12 May 1991* (Innsbrucker Beiträge zur Kulturwissenschaft / Anglistische Reihe 4), 155–166. Innsbruck: Verlag des Institutes für Sprachwissenschaft.

Kunz, Kerstin, Ekaterina Lapshinova-Koltunski, José M. Martínez Martínez, Katrin Menzel & Erich Steiner. 2021. *GECCo - German-English Contrasts in Cohesion: Insights from Corpus-Based Studies of Languages, Registers and Modes* (Trends in Linguistics. Studies and Monographs 355). De Gruyter Mouton.

Rohdenburg, Günter. 1990. Aspekte einer Vergleichenden Typologie des Englischen und Deutschen: Kritische Anmerkungen zu einem Buch von John A. Hawkins. *Kontrastive Linguistik. Forum Angewandte Linguistik* (19). 133–152.

Ström Herold, Jenny & Magnus Levin. 2018. English Supplementive ing-Clauses and their German and Swedish Correspondences. *BeLLS* 9(1). 115–138.

The impact of the intended reader on language complexity: a contrastive view of supplementive participial clauses in children's fiction

Markéta Malá (Charles University, Prague)

In fiction for children, both the content and the language are "adjusted to readers' comprehension and reading abilities" (Puurtinen 1998: 2), with language complexity being reduced to enhance readability. To what extent a by what means the appropriate level of complexity is retained in translation appears to be highly dependent on the languages in question (Barbieri Durão & Kloeppel 2018). We rely on English (source) and Czech (target) texts, exploring the role of three potential factors contributing to language complexity: the reader (children vs. adults), the writer and the language. To investigate their impact, we analyse books of three English authors who wrote both for children and for adults, and whose works have been translated into Czech (J.K. Rowling, R. Dahl, N. Gaiman).

Even though the corpus is (still) small (650 and 575 thousand words in the English and Czech sub-corpora, respectively), it makes it possible to apply contrastive corpus stylistic methods (Toolan 2018), supported by detailed analysis of concordance lines, using *cqpweb* and *KonText* tools.

A global point of view of the texts highlights, for instance, lower lexical diversity of children's books in both languages, compared to fiction for adults. What we focus on, though, is rather linguistic complexity at the local level, i.e. the structure complexity of individual linguistic features (Bulté & Housen 2012). Comparing English and Czech, the inflectional character of the latter is particularly prominent in its reliance on finite verb predicates and limited use of non-finite clauses. Participial supplementive clauses, e.g. *Breathing very fast, he turned slowly back to the mirror*, have no congruent counterpart in Czech (Malá & Šaldová 2015). The processing complexity of English supplementive clauses is accounted for by their "implicit and somewhat ill-defined relationship with the main clause" (Biber et al. 1999: 782- 3, cf. Ström Herold & Levin 2018). In sentence-initial participial clauses, moreover, the identification of the unexpressed subject may cause problems. Despite this, sentence-initial supplementive clauses occur in children's fiction, albeit less frequently than in adult's books. A more detailed view shows that while Dahl and Gaiman rarely use them in books for either type of readers, both Rowling's detective novels and children's books abound in these structures (8.6 and 5.4 per 1000 sentences, respectively). The reader, however, constitutes a factor leading to the adjustment of complexity of supplementive clauses in her novels: while in *The Silkworm* sentence-initial participial clauses tend to be long (mean 7.6 words) and internally complex, comprising subordinate clauses, in *Harry Potter and the Philosopher's Stone* they are shorter (mean 5.5 words) and structurally simple.

The Czech translations of initial supplementive clauses may be divided into more and less explicit ones: the former comprise finite clauses (co-/subordinate), which are easier to process due to the verbal categories expressed overtly by the finite verb in Czech, subordinators, and in some cases the shift of the proper-noun subject to the sentence-initial clause. Less explicit counterparts include adverbials expressed by adverb or prepositional phrases. The tendency towards reducing structure complexity in children's fiction is reflected in the preference for the more explicit counterparts in the Czech translations of the children's books. This suggests that regard to the reader can influence the degree of complexity both in the source and translated texts.

References

Barbieri Durão, A. B. de Amorim & P. R. Kloeppel. 2018. Children's literature parallel corpora: a hybrid experimental model to evaluate transfers of language complexity via linguistic transcoding. *Ilha do Desterro* 71(1): 27-51.

Biber, D. et al. 1999. *Longman Grammar of Spoken and Written English.* Harlow: Longman.

Bulté, B. & A. Housen. 2012. Defining and operationalising L2 complexity. In Housen, A., F. Kuiken & I. Vedder (eds) *Dimensions of L2 Performance and Proficiency. Complexity, Accuracy and Fluency in SLA.* Amsterdam: John Benjamins. 21-46.

Malá, M. & P. Šaldová. 2015. English non-finite participial clauses as seen through their Czech counterparts. *Nordic Journal of English Studies* 14(1): 232-257.

Puurtinen, T. 1998. Syntax, readability and ideology in children's literature. *Meta* 43(4): 524- 533.

Ström Herold, J. & M. Levin. 2018. English supplementive *ing*-clauses and their German and Swedish correspondences. In Oksefjell Ebeling, S. & H. Hasselgård (eds) *Corpora et comparatio linguarum: Textual and Contextual Perspectives*, *BeLLS* 9(1): 115-138.

Toolan, M. 2018. How children's literature is translated: Suggestions for stylistic research using parallel corpora. *Ilha do Desterro* 71(1): 151-167.

https://cqpweb.lancs.ac.uk/ (Accessed April 2, 2022)

InterCorp, v. 14. Institute of the Czech National Corpus, Faculty of Arts, Charles University, Prague. 2022. Available from: http://www.korpus.cz (Accessed April 2, 2022)

Structural and semantic features of adjectives across languages and registers

Signe Oksefjell Ebeling (University of Oslo)

This paper investigates the use of adjectives in comparable English and Norwegian fictional texts. The aim is to shed light on how the two languages make use of adjectives to describe fictional worlds. In a previous contrastive study of 'be' verbs in Czech, English and Norwegian, it was found that the use of adjectives in predicative function, i.e. in the NP BÝT/BE/VÆRE + ADJ pattern, is a more defining feature of English fiction than of the other two languages in terms of frequency (Čermáková et al. Submitted). Moreover, even if the analysis suggested that the three languages resort to similar strategies to describe fictional subjects by means of (predicative) adjectives, some differences were noted in the semantic quality of the adjectives used in the languages compared.

These findings triggered some questions for further research, including: (i) Does Czech and Norwegian fiction make more frequent use of attributive adjectives to convey the same message as English does with predicative adjectives? and (ii) Is English fiction generally more concerned with characterizing or describing the Subject in terms of (specific sets of) adjectives? In addition to addressing these questions from an English-Norwegian perspective, the current study will take the hybrid nature of fiction into account by separately investigating the use of adjectives in passages representing fictional dialogue vs. narrative. This added complexity of operating with two registers (dialogue and narrative) instead of one (fiction) has proved to be an important one when investigating the language of fiction, as previous studies have pointed to linguistic differences between the two sub-registers in English (e.g. Egbert & Mahlberg 2020; Ebeling & Hasselgård 2020), also cross-linguistically (e.g. Ebeling & Ebeling 2020).

Preliminary observations of material from the general fiction part of the English-Norwegian Parallel Corpus suggest that there are some differences regarding attributive vs. predicative use, both across the languages and registers. Contrary to expectations, English seems to prefer the attribute function in both dialogue (57%) and narrative (73%), e.g. example (1) from the narrative sub-corpus. Similarly, Norwegian prefers the attributive function in narrative (61%), while there is more of an equal distribution between attributive (44%) and predicative (43%) function in Norwegian dialogue, e.g. (2).[1]

(1) His *big* feet were on the table. (BO1n)
(2) "Margaret er *snill*," *...* (BV2d)
Margaret is *nice*, ... (BV2dT)

This initial scrutiny of a small sample suggests that the short answer to the first research question is "No", and that other, more complex, factors may be at play to account for the previously noted discrepancy between the languages, e.g. Norwegian may make use of other verbs than the prototypical 'be' verb (VÆRE) to express this relationship between subject and predicative adjective, or there may be a general tendency for English to use more adjectives in fiction overall. These issues will be investigated on the basis of a larger sample and will in turn prepare the ground for a semantic analysis of the adjectives that will address the second research question in particular.

[1] In the remaining 13%, the adjective has been classified as neither attributive nor predicative, e.g. when it functions as head of a noun phrase as in …*dra den vesle med seg* … 'drag the little [one] with her'.

References

Čermáková, Anna, Jarle Ebeling and Signe Oksefjell Ebeling. Submitted. 'Be' verbs in a contrastive perspective: The case of BÝT, BE and VÆRE.

Ebeling, Signe Oksefjell and Jarle Ebeling. 2020. Dialogue vs. Narrative in fiction. A cross-linguistic comparison. *Languages in Contrast* 20:2, 288–313.

Ebeling, Signe Oksefjell and Hilde Hasselgård. 2020. Intensification in dialogue vs. narrative in a corpus of present-day English fiction. In E. Jonsson & T. Larsson (eds), *Voices past and Present – Studies of Involved, Speech-Related and Spoken Texts. In Honor of Merja Kytö*. Amsterdam: Benjamins. 301–316.

Egbert, Jesse and Michaela Mahlberg. 2020. Fiction – one register or two? Speech and narration in novels. *Register Studies* 2:1, 72–101.

Mood and modality: The Spanish subjunctive and its English counterpart(s)

Rosa Rabadán & Noelia Ramón
(University of León)

Complexity refers to the internal structuring of linguistic units or systems regarding the number and variety of their components and relationships (Dahl 2004, Miestamo 2008, Housen et al. 2019). Complexity differences between languages in morphology tend to be compensated elsewhere, for example, in syntax. These cross-linguistic complexity differences have served traditionally as a yardstick for contrasting languages. This paper examines mood and modality as one of such differences.

Modality is a meaning notion concerned with the speaker's attitude towards the (non-)factuality of the situation. Mood is a category of grammar dealing with the grammatical marking of modality in the verbal system (Palmer 2001). As a Romance language, Spanish marks mood in the inflectional system of the verb. English, by contrast, as a Germanic language, displays only residual mood inflections and relies on the unmarked forms of the verb and lexical resources to convey modal meanings, e. g. *Fache could not imagine anyone risking a stunt like this. (FBD1E.s36)/ A Fache le* resultaba inconcebible *que alguien se arriesgara a dar un salto como aquel. (FBD1S.s34)*, where the non-finite clause turns into an inflected subjunctive in Spanish.

Both English and Spanish feature a three-mood system: indicative, imperative and subjunctive (Bolinger 1970). The indicative mood is the unmarked one in the two languages, while the imperative is devoted exclusively to conveying directives. However, the Spanish subjunctive covers most modalized situations, such as doubt, volition, necessity, likelihood, and marking non-factive temporality (Jiménez Juliá 1989, Bosque 1990, González Calvo 1995). Spanish also has the possibility of using lexical modals and subordination (Fuentes Rodríguez 1991). 'Subjunctive-poor' English uses different resources to account for modal meanings, including lexical modals, certain types of subordination, verb forms in the indicative mood, and modal auxiliaries (Huddleston & Pullum 2002: 173-175).

This paper explores how English and Spanish modalized forms relate cross-linguistically as part of a broader project on verb contrast. Data come from the English-Spanish parallel corpus P-ACTRES 2.0, a bidirectional English-Spanish corpus consisting of original texts in one language and their translation into the other. P-ACTRES 2.0 contains about 6 million words considering both directions. We have used the English into Spanish subcorpus for this paper, totalling 4,296,733 words. Concerning register, the corpus features fiction and non-fiction materials.

We have used a back-translation procedure: Starting from the Spanish translations, we have traced the origins of the subjunctive imperfect forms (first- and third-person singular) to English. As P-ACTRES 2.0 is PoS tagged and the tag set for English-Spanish does not provide mood distinction, we started by querying the *-se* and *-ra* endings of the imperfect subjunctive inflections. However, we needed additional discriminatory features to obtain more focused results. We added the vowel alternation that marks the first conjugation from the other two, and our final input was *-ara/-era* and *-ase/-ese*. Still, the results were sampled and manually filtered to discard forms unrelated to our query.

In our pilot study, we have analyzed 1,175 concordances featuring Spanish subjunctive forms in translation. The English source data indicate that the main triggers for these subjunctive solutions are: indicative past tenses (30%), lexical modals (23.2%), non-finite clauses (13.07%), conditional constructions (11.8%), modal auxiliaries

(10.5%), and subjunctive mood forms (2.2%). The past tenses tend to appear in subordinate clauses where the matrix clause encodes modal notions such as doubt, volition, necessity, and likelihood or mark non-factuality. Additionally, our data have yielded several instances (7.55%) where the English original does not trigger the subjunctive in Spanish but results from translation strategies. According to these results, the generalized idea that the role of the Spanish subjunctive mood is performed in English by modal auxiliaries is not supported by evidence (Rabadán 2006, 2007).

References

Bolinger, D. 1970. Modes of Modality in Spanish and English. *Romance Philology* 23(4): 572-580.

Bosque, I. 1990. Las bases gramaticales de la alternancia modal. Repaso y balance. In I. Bosque (ed.) *Indicativo y subjuntivo*. Madrid: Taurus. 13-65.

Dahl, Ö. 2004. *The Growth and Maintenance of Linguistic Complexity*. John Benjamins.

Fuentes Rodríguez, C. 1991. Adverbios de modalidad. *Verba* 18: 275-321.

González Calvo, J. M. 1995. Sobre el modo verbal en español. *Anuario de estudios filológicos* 18: 177-204.

Housen, A., B. De Clercq, F. Kuiken and I. Vedder. 2019. Multiple approaches to complexity in second language research. *Second Language Research* 35(1): 3-21.

Huddleston, R. and G.K. Pullum. 2002. The Cambridge Grammar of the English Language.

Jiménez Juliá, T. 1989. Modalidad, modo verbal y modus clausal en español. *Verba* 16: 175-214.

Miestamo, M. 2008. Grammatical complexity in cross-linguistic perspective. In Miestamo, M. K. Sinnemäki and F. Karlsson (eds.) 2008. *Language Complexity. Typology, contact, change*. John Benjamins. 23-41.

Palmer, F. R. 2001. *Mood and Modality* (2nd ed.). Cambridge: Cambridge University Press.

Rabadán, R. 2006. Modality and modal verbs in contrast: mapping out a translation(ally) relevant approach English-Spanish. *Languages in Contrast* 6(2): 261-306.

Rabadán, R. 2007. Translating the 'predictive' and 'hypothetical' meanings English-Spanish. *Meta* 52(3): 484-502.

Stylistic repetition in non-fiction writing:
Contrastive and translational perspectives

Jukka Tyrkkö (Linnaeus University)

Stylistic repetition, or intentional lexical parallelism, is one of the classical rhetorical devices. In this pilot study, particular attention is given to anaphoric and epiphoric repetition localised to short segments within a text, a practice that increases salience and highlights the purposeful nature of the repetition. Whilst stylistic repetition in literary texts has been discussed extensively in both corpus stylistics (Starcke 2006, Mahlberg 2013) and translation studies (e.g. Boase-Beier 1994, Edwards 1997, Čermáková 2015, Klinger 2019), few scholars to date have examined the prevalence of repetitions in non-fiction texts, nor what translators do with them. It is widely acknowledged that the reduction of repetition is a common strategy in translating (Toury 1991, Ben-Ari 1998, Chesterman 2006, Laviosa 2009), the observation has been made in reference to lexical tautology, rather than to repetitions of longer stylistically motivated sequences. Studies contrasting the use of stylistic repetition across languages are likewise scarce (see Pacheco 1992, Al-Mukharriq 1993, Niu & Hong 2010). Thus, this gap in research prompts questions such as how common is stylistic repetition in non-fiction writing, and do languages differ when it comes to the intentional use of repetitions?

The Linnaeus University English-German-Swedish (LEGS) three-way parallel corpus currently comprises 30 texts representing biographies, guide books, popular science, historical accounts, and other similar non-literary genres (see, e.g., Ström Herold & Levin 2018). In the example below from LEGS, repetition in the English source text (a) is replicated fairly closely in the Swedish target text (b) but largely ignored by the German translator (c).

a) Where could you turn for safety?
Where could you turn for comfort?
Where could you turn for meaning?

b) Vart vänder du dig för att få trygghet?
Vart vänder du dig för tröst?
Vart vänder du dig för att finna mening?

c) Wo könnten Sie Zuflucht finden?
Oder Trost?
Was könnten Sie tun, um sich zu orientieren?

First, using LEGS as primary data, language-specific baseline data are established for repetitive sequences, starting with a definitional exploration of different thresholds of n-gram lengths and windows of repetition. The n-grams were retrieved with a custom pattern-analysis tool (to be made available at the workshop), classified into three types (*functional*, *nominal*, or *hybrid*), and contrasted across languages, comparing original and translated texts in each language. Second, for each repetition in the source texts, the presence or lack of a corresponding repetitions in the target text is examined. Using the repetition type, the length of the n-gram, and the mean distance between the n-grams in a repetition cluster as predictors, the likelihood of translators reproducing repetitions encountered in the source texts is analysed for each language pair. The statistical analysis

is followed up with a qualitative examination of the translation strategies with particular reference to the availability of equivalent structures in the target language.

References

Al-Mukharriq, Hayfa'. 1993. Repetition as an Effective Rhetorical Device in Arabic and English Argumentative and Expository Texts. Unpublished PhD thesis. University of Glasgow. Available online at <https://theses.gla.ac.uk/76301/1/13834033.pdf>

Ben-Ari, Nitsa. 1998. The Ambivalent Case of Repetitions in Literary Translation. Avoiding Repetitions: A "Universal" of Translation? *Meta: Journal Des Traducteurs* 43:1. 1-75.

Boase-Beier, Jean. 1994. Translating repetition. *Journal of European Studies*, 24:4-96. (1994:Dec.) p.403-409.

Chesterman, Andrew. 2006. Interpreting the meaning of translation. SKY Journal 19:1. Available online at <http://www.linguistics.fi/julkaisut/SKY2006_1/1FK60.1.1.CHESTERMAN.pdf>

Čermáková, Anna. 2015. Repetition in John Irving's novel A Widow for One Year. A corpus stylistics approach to literary translation. *International Journal of Corpus Linguistics* Vol 20:3. 355–377

Edwards, Michael. 1997. Translation and repetition. *Translation and Literature*, 6:1. 48-65.

Klinger, Susanne. 2019. Repetition. Translating the interplay between its linguistic form and its literary function. *Babel* 65:2. 316-332.

Laviosa, S. 2009. Universals. In Baker, Mona (ed.) 2009. *Encyclopedia of Translation Studies*. London: Routledge. 306–311.

Linnaeus University English-German-Swedish corpus (LEGS). 201?-. A parallel corpus compiled at Linnaeus University by Magnus Levin and Jenny Ström-Herold. <https://lnu.se/en/research/searchresearch/the-linnaeus-university-english-german-swedish-corpus-legs/>

Macken, Lieve. 2010. In search of recurrent units of translation. In Daelemans Walter & Véronique Hoste (eds.) *Evaluation of Translation Technology*. Antwerp: ASP Editions. 195–212.

Mahlberg, Michaela. 2013. *Corpus Stylistics and Dickens's Fiction*. London: Routledge.

Niu, Guiling & Huaqing Hong. 2010. Repetition patterns of rhetoric features in English and Chinese advertisements: A corpus-based contrastive study. In Xiao, R. (Ed.), *Using Corpora in Contrastive and Translation Studies. Translation Studies.* London: Cambridge Scholars Publishing. 433-456.

Pacheco, Jorge Arturo Quesada. 1992. Functions of repetition in two Western languages: English and Spanish. Filología y Lingüística XVIII:1. 163–176.

Starcke, B. 2006. The phraseology of Jane Austen's Persuasion: Phraseological units as carriers of meaning. *ICAME Journal*, 30, 87–104.

Ström Herold, Jenny & Magnus Levin. 2018. English supplementive ing-clauses and their German and Swedish correspondences. *Bergen Language and Linguistics Studies*. 9 (1). 115-138.

Toury, Gideon. 1991. What are descriptive studies into translation likely to yield apart from isolated descriptions? *Translation Studies: the state of the art. Proceedings of the First James S Holmes Symposium on Translation Studies*. 179-192.

Papers

A contrastive study of -ish in English and Swedish

Karin Aijmer (University of Gothenburg)
karin.aijmer@sprak.gu.se

The aim of my presentation is to compare the frequency and use of -ish in English and Swedish based comparable corpora from the perspective of pragmatics and pragmatic borrowing (Andersen 2020, 2021). -ish is not a Swedish suffix or word but is borrowed from English suffix meaning approximation (exemplified by tallish, bluish). It is now also used as a a pragmatic marker with a hedging or qualifying function where it can be compared with sort of (Kuzmack 2007, Peirce 2014,2015). It is spreading to other languages such as Norwegian (Nilssen and Kinn 2017) and Dutch (Norde 2009). On the basis of a comparison of the use and function of -ish both similarities and differences can be investigated. The data gives rise to several research questions which concern the analysis of – ish and the conditions under which it is borrowed into English. How can we distinguish between the free and bound uses of -ish? How are they related to different functions? Which categories of -ish are most likely to be borrowed? Do English speakers and Swedish speakers use -ish in the same way and for the same functions?

The cross-linguistic analysis of -ish is based on the occurrences in blogs retrieved from the Birmingham Blog Corpus https://wse1.webcorp.org.uk/home/blogs.html and from a Swedish blog corpus (included in the Korp Corpus https://spraakbanken.gu.se/korp/.). In both corpora Ish is used as a suffix attached to many types of words and phrases and it can be used on its own. In both English and Swedish ish is frequent after numerals to indicate an approximate time or place (often with a collocating item with the meaning 'about'.

(1) on Sunday we arrived about 9 ish
(2) Är med svintidigt , 07.20 ish
('Is there really early, about 7.20')

The following examples illustrate how it is used on its own as a pragmatic marker which can be paraphrased as 'sort of':

(3) I'm working on another darker YA and a thriller that's somewhere between YA and adult (kind of like the Power Rangers meets Da Vinci Code) (Ish) (which is totally a thing)

(4) You're right. Ish ignoring things is hardly EVER good strategy.

As a pragmatic marker it is associated with the expression of speaker's attitudes The same usage is illustrated in Swedish. As shown in example (6) it collocates with 'typ' with the same function.

(5) Terobi Jag tror bestämt vi är på samma plats . Ish .
(I think definitely we are in the same place. Ish.)

(6) Ja – det här är alltså det bästa som har hänt internet sen … slajsat bröd. Typ . Ish . Julbockens sång - Ted Åström
(Yes- This is consequently the best that has happened to internet since… sliced bread. Type. Ish.)

In examples (x) – (x) ish modifies a whole proposition as a pragmatic marker which can be compared to English sort of/kind of and the Swedish typ 'type'. The preliminary results indicate that ish is borrowed both as a suffix and a pragmatic marker but in different contexts. In Swedish -isc is also used as before the word it modifies as in (7):

(7) Jag kan komma förbi ish 11.
('I can come by around 11').

References

Andersen, G. 2020. Three cases of phraseological borrowing: A comparative study of as I, oh wait and the ever construction in the Scandinavian languages. Ampersand 7(1-9).

Andersen, G. 2021. What governs speakers' choices of borrowed vs. domestic variants of discourse-pragmatic variables? In Peterson, E. , Hiltunen, T. and J. Kern (eds), Discourse-Pragmatic Variation and Change in English. New methods and insights. 251-271. Cambridge: Cambridge University Press.

Kuzmack, S. 2007. Ish: How a suffix became a word. Ms, University of Chicago.

Nilssen, S. and T. Kinn 2017. A chameleon abroad: English -ish and ish used in Norwegian. Maal and minne 2 (123-143)

Norde, M. 2009. Degrammaticalization. Oxford: Oxford University Press.

Pierce, M. 2014. The further degrammaticalization of -ish. American Speech 89(1): 115-118.

Pierce, M. 2015. More on -ish. American Speech 89(1): 115-118.

The *Hansard Corpus*: Semantics and scaffolding

Marc Alexander, Fraser Dallachy, Ewan Hannaford (University of Glasgow)

The *Hansard Corpus* of speeches in the British Parliament (initially 1803-2005 at https://www.english-corpora.org/hansard/) has recently been updated until 2020 by the *Hansard at Huddersfield* team (https://hansard.hud.ac.uk/), while the Glasgow team have also been working on research on the frequency distribution of semantic categories. This paper will report work in progress on the highest-frequency categories which appear when using corpora tagged with the Historical Thesaurus Semantic Tagger (HTST; see Piao *et al* 2017 and Alexander *et al* 2015), which uses the *Historical Thesaurus of English* (Kay *et al* 2022) as its underlying data. The paper has a focus on the *Hansard Corpus* (Alexander and Davies 2015), but also uses the *Semantic EEBO* corpus (the HTST-tagged version of Early English Books Online).

Our work on a study of these high-frequency semantic categories is firstly an exercise in corpus linguistics and secondly aims to help improve a future version of the tagger. A frequency analysis of HTST tags shows a standard Zipfian pattern, much like any corpus, but the characteristics of this distribution are different to that of standard lexical distribution. While high-frequency lexical items are often grammatical items, our initial results show that high-frequency HTST categories often serve functions on the margins between content and structure. For example, they can act as:

- markers of topic, genre, or stance,
- organisational units regarding the semantic structure of the discourse and discourse relationships, and
- clusters of widely-lexicalised concepts which often contain a spread of low-frequency words.

In the paper, we will discuss short examples of these and their contribution to the distribution of semantic categories in discourse. These include adverbs which represent the traditional relationships of place, time, circumstance, manner, degree, and cause (captured within the *Relative Properties* tag), common discourse concerns which are widely lexicalised and when clustered show the containing concept's high frequency (such as *Number*), and categories whose relative internal distribution expresses information about the text as a whole (such as *Time*, which is a highly-common concept but is realised through different combinations of particular and indefinite temporal reference based on factors including authorship, genre, stance, and so forth). Our future work involves a systematic categorisation based on cross-textual comparisons, as well as work on the tagging itself, to investigate how these semantic domains contribute to the 'normal' discursive frames that scaffold and shape texts.

References

Alexander, Marc & Mark Davies. 2015. *The Hansard Corpus, 1803-2005.* Available online at http://www.english-corpora.org/hansard.

Alexander, Marc, Fraser Dallachy, Scott Piao, Alistair Baron & Paul Rayson. 2015. Metaphor, popular science, and semantic tagging: Distant reading with the *Historical Thesaurus of English*. *Digital Scholarship in the Humanities* 30(s1). 16–i27, doi: 10.1093/llc/fqv045

Hansard at Huddersfield Project. 2018–. *Hansard at Huddersfield*. University of Huddersfield. Available online at: https://hansard.hud.ac.uk.

Kay, Christian, Marc Alexander, Fraser Dallachy, Jane Roberts, Michael Samuels, and Irené Wotherspoon (eds.). 2022. *The Historical Thesaurus of English* (2nd edn., version 5.0). University of Glasgow. https://ht.ac.uk/.

Piao, Scott, Fraser Dallachy, Alistair Baron, Jane Demmen, Stephen Wattam, Philip Durkin, James McCracken, Paul Rayson & Marc Alexander. 2017. A time-sensitive *Historical Thesaurus*-based semantic tagger for deep semantic annotation. *Computer Speech and Language* 46. 113-135. doi: 10.1016/j.csl.2017.04.010.

KWIC Patterns: A new normal for displaying, ordering, and interpreting concordance line results

Laurence Anthony (Waseda University)
anthony@waseda.jp

Key-Word-In Context (KWIC) concordance lines are one of the oldest and most common tools used by corpus linguists. Generating and interpreting KWIC concordance lines is also a key strategy used in a data-driven learning (DDL) approach to second and foreign language teaching. The fundamental way that KWIC concordance lines are displayed has remained almost unchanged for over 60 years. After the researcher, teacher or learner makes a search query, the corpus analysis software finds hits in the target corpus and displays these on separate lines with by a fixed number of surrounding words or characters to provide context. An unordered set of concordances lines is almost impossible to interpret except when the number of results is very small. Thus, most software tools offer an option to order the lines based on word positions to the left or right of the query hit. Intriguingly, this sort option has invariable been based on an alphabetical ordering of the words in the target positions, which naturally prioritizes concordance lines with target words starting with "a" (for English) regardless of the linguistic importance of such lines. Because of this design choice, interpreting concordance line inevitable requires users to scroll through huge numbers of concordance lines while trying to notice salient patterns in the results. Some tools have an option to show a random subset of results or every 'nth' result, which partially alleviates the data-overload issue. However, the fundamental problem associated with concordance line ordering remains unsolved and perhaps even unnoticed in the field at large.

In this paper, I will introduce a 'new normal' for displaying, ordering, and interpreting concordance lines based on a concept called "KWIC Patterns". KWIC patterns are similar in concept to word 'clusters'. However, rather than being continuous multi-word units (MWUs) built around a search query, KWIC patterns are continuous or non-contiguous MWUs that are determined by the word positions set in the concordance tool, e.g., MWUs built on the words that appear one word to the left (1L) and one word to the right (1R) of the search query word(s). Crucially, these KWIC patterns can be ranked by frequency of occurrence, allowing the KWIC concordance lines that contain such patterns to also be ranked accordingly. Results show that KWIC pattern ordering of concordance lines dramatically improves the ease-of-interpretation of concordance displays. Salient patterns always appear first in the display and the relative importance of each salient pattern can be easily evaluated by simply scrolling down the list of results. KWIC pattern ordering also dramatically improves the learners' experience in a DDL classroom as it is no longer necessary to navigate through long lists of alphabetically ordered concordance lines to find patterns of importance. KWIC pattern ordering has been added to the AntConc (Anthony, 2021) corpus toolkit in its latest release (4.0), making the functionality available for both left-to-right and right-to-left languages, as well as ideographic languages such as Japanese and Chinese.

The FUTURE of World Englishes: will versus BE going to in the International Corpus of English

Axel Bohmann (University of Freiburg)
axel@bohmann.de

The choice between will and BE going to (BGT) for marking future-time reference (FTR) is a well-established alternation in English (e.g. Denis & Tagliamonte 2018). Historically, both variants derive from processes of grammaticalization. The development of will from a modal of volition/intention to a future marker can be traced back as far as Old English, whereas the grammaticalization of BGT sets in towards the end of the Middle English period (Fischer 1992: 265). Competition between these two forms since then is characterized by a gradual increase in BGT, which however has not gained dominance over will as yet (Szmrecsanyi 2003: 296). In present-day English, the alternation is subject to stylistic, regional, as well as syntactic constraints. However, systematic consideration of will versus BGT in World Englishes is still lacking.

The present study addresses this empirical gap by considering FTR constructions with will and BGT in eight corpora of the International Corpus of English (ICE) project (Greenbaum & Nelson 1996), representing American, Canadian, New Zealand, Indian, Sri Lankan, Singaporean, Hong Kong and Jamaican Standard English. More than 27,000 such constructions are extracted from the part-of-speech tagged versions of these corpora and subjected to multivariate regression analysis with the following predictors (following Denis & Tagliamonte 2017): subject (first-person singular vs. other animate vs. inanimate), polarity, clause type (main versus subordinate clause), sentence type (declarative versus interrogative), presence of a temporal adverbial, and text category. In addition, the national variety each token is extracted is included both as a main effect and in interaction terms with the other predictors.

Results are generally in line with the previous literature: texts in the written modality and the presence of temporal adverbials favor will, whereas interrogatives and first-person subjects favor BGT. Beyond these effects, however, there is also pronounced cross-regional differentiation in the FTR alternation. Particularly, Asian varieties show a strong favoring of will compared to all other varieties. The difference is not simply one of overall preference, but affects the strength of individual predictors, as significant interaction terms for all main effects with the predictor variety demonstrate.

The findings are descriptively valuable since they offer a previously lacking account of global differentiation in the English FTR system. At the theoretical level, they are relevant for improving our understanding of relationships among varieties in the World System of Englishes.

References

Denis, Derek & Sali Tagliamonte. 2018. The changing FUTURE: competition, specialization and reorganization in the contemporary English future temporal reference system. English Language and Linguistics 22(3), 403-430.

Greenbaum, Sidney & Gerard Nelson. The International Corpus of English (ICE) project. World Englishes 15(1), 3-15.

Fischer, Olga. Syntax. In N. Blake (ed.) The Cambridge History of the English Language. Vol. 2: 1066–1476. Cambridge: Cambridge University Press, 207–408.

Szmrecsanyi, Benedikt. 2003. 'Be going to' versus 'will/shall': Does syntax matter? Journal of English Linguistics 31(4), 295–323.

Lexical sophistication in spoken English: Lex Complexity Tool and the Spoken BNC2014 wordlist

Raffaella Bottini & Vaclav Brezina
(Lancaster University)

Lexical sophistication is a complex construct, which is based on the identification of the proportion of low-frequency lexical items in texts and transcripts of speech. Lexical sophistication measures depend on (i) the lexical unit to identify words, (ii) the reference corpus to measure frequency, and (iii) the method that rates frequency in a target text (Kyle, 2019). Several tools for the automatic computation of lexical complexity have been developed (e.g. TAALES, Kyle & Crossley, 2015; Coh-Metrics, Graesser et al., 2004; Lexical Complexity Analyzer, Lu, 2012) and their performance varies in terms of functionalities, transparency, accessibility, and customisation (cf. Kyle, 2019). The majority of existing tools have been developed to analyse written language since they are based on written reference corpora and rarely include datasets of spoken language, often consisting of scripted speech. This study aims to contribute to corpus-based vocabulary research presenting Lex Complexity Tool, an automated tool based on a Python code which was developed for the computation of a wide range of lexical complexity scores. The tool includes existing and new indices, as well as a new wordlist extracted from the Spoken BNC2014 (Love et al., 2017), a large reference corpus of spontaneous L1 English speech which makes Lex Complexity Tool particularly suitable for the lexical analysis of L1 and L2 spoken corpora. The value of this methodological approach is demonstrated through a case study combining quantitative and qualitative analysis of data from the 4.2-million-word Trinity Lancaster Corpus (TLC; Gablasova et al., 2019). The TLC consists of transcripts of learners' spoken performance based on the Graded Examination in Spoken English (GESE) which is a high-stakes exam of L2 English developed and administered by Trinity College London, a large international examination board. The results show differences across lexical sophistication indices in terms of their sensitivity and interpretability. Band-based metrics of sophistication which distinguish different word classes could be a preliminary straightforward method for assessing vocabulary, accessible to non-experts, such as teachers, learners, and language testers. Mean-frequency indices, especially based on content words, provide a more fine-grained picture of lexical sophistication; their non-transformed values might be preferred to their logarithmic transformations since they allow a more precise comparison of scores across texts. Methodological and practical implications for corpus-based studies on spoken language are discussed.

References

Gablasova, D., Brezina, V., & McEnery, T. (2019). The Trinity Lancaster Corpus: Development, description and application. International Journal of Learner Corpus Research, 5(2), 126–158.

Graesser, A., McNamara, C., Louwerse, D., & Cai, S. (2004). Coh-Metrix: Analysis of text on cohesion and language. Behavior Research Methods, Instruments, & Computers, 36(2), 193–202.

Kyle, K. (2019). Measuring lexical richness. In S. Webb (Ed.), The Routledge handbook of vocabulary studies (pp. 454–476). Routledge.

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. TESOL Quarterly, 49(4), 757–786.

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. International Journal of Corpus Linguistics, 22(3), 319–344.
Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. The Modern Language Journal, 96(2), 190–208.

While the Beacon-fire blazed its brightest, the two women shrieked their loudest: On the Superlative Object Construction (SOC)

Tamara Bouso

Little attention has been paid to the Superlative Object Construction (SOC), as in the examples in (1)-(3).

(1) She **worked** *her hardest*.
(2) The two women **shrieked** *their loudest*.
(3) The Beacon-fire **blazed** *its brightest*.

The SOC consists of an intransitive verb of manner of action (*work*, *shriek*, *blaze*, etc.) followed by a non-prototypical type of object that takes the form of a nominalised adjective, preceded by a possessive, and inflected for the superlative degree (*her hardest*, *their loudest*, *their brightest*, etc.). Its overall constructional meaning is one of intensification, namely, to express the highest level on a scale of a property that qualifies the matter of the action denoted by the verb.

To the best of my knowledge, the historical grammarians Jespersen (1909-1949) and Poutsma (1914-1929) are the only ones who do touch on the SOC, and they do so in passing, relying also on what seem to be the prototypical examples of the construction (1-3). This empirical evidence is not sufficient to provide a detailed analysis of the form, function, frequency, and distribution of the SOC in Present Day English (PDE). This is the aim of this paper: to offer a full-fledged characterization of the modern SOC from the perspective of Construction Grammar (Goldberg 1995, 2006, Hilpert 2013, 2014/2019), and on the basis of naturally occurring data from the *Corpus of Contemporary American English* (COCA, Davies 2008). More concretely, the research questions that will be addressed are the following:

i) What are the characteristic features of the modern SOC?
ii) How frequent and productive is the SOC in PDE?
iii) Does register play a role in the use of the construction? If so, how?

It will be argued that the SOC qualifies as an intensifying comparative construction. Like other analogous structures involving non-prototypical object types (Bouso 2021), the SOC counts as a traditional Goldbergian construction. It has unusual syntax (see 4), and it also lacks compositionality as inflectional superlatives most naturally express comparison rather than intensification (Huddleston and Pullum et al. 2002, 1165fn41).

(4) FORM: $SUBJ_i$ [$V_{TRANS\ /\ INTRANS}$ OBJ_i]. Where OBJ_i = $(POSS)_i$ NP-*est* $_{ADJ}$ ↔ MEANING: '$Agent_i$ cause OBJ_i become expressed in its highest degree by doing V' {intensification}

Similarly to other emphatic comparative constructions, like the [*más feo que* X] ('very ugly') construction (Ivorra Ordines 2021), despite being low frequent and showing a considerable high number of entrenched lexicalized units (*do [X] best*, *try [X] best*, *look [X] best*, etc.), the SOC is relatively productive. To be more specific, the SOC can be treated as a polysemous construction structured around two core meanings: (i) *to be in one's best state or condition*, featuring (copular) stative verbs such as *feel* and *look*, and (ii) to *do X at one's highest standard or levels*. This second sense involves (in)transitive verbs of manner of action of various kinds (*do*, *try*, *play*, *work*, *smile*, *roar*, etc.), and is primarily a

characteristic of informal registers — blogs, magazines, and journalist discourse in general —, where the SOC can be easily accommodated to serve emotive, phatic, and conative functions.

References

Bouso, Tamara. 2021. *Changes in Argument Structure. The Transitivizing Reaction Object Construction*. Bern: Peter Lang.

Davies, Mark. (2008-) The Corpus of Contemporary American English (COCA). Available online at https://www.english-corpora.org/coca/.

Goldberg, Adele. E. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago, IL: The University of Chicago Press.

Goldberg, Adele E. 2006. *Constructions at Work: The Nature of Generalization in Language*. Oxford: Oxford University Press.

Hilpert, Martin. 2014/2019. *Construction Grammar and its Application to English*. Edinburgh: Edinburgh University Press.

Hilpert, Martin. 2013. *Constructional Change in English: Developments in Allomorphy, Word Formation, and Syntax*. Cambridge: Cambridge University Press.

Huddleston, Rodney, and Geoffrey K. Pullum, et al. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Ivorra Ordines, Pedro. 2021. "Comparative constructional idioms: A corpus-based study of the creativity of the [*más feo que* X] construction." In *Productive Patterns in Phraseology and Construction Grammar: A Multilingual Approach*, edited by Carmen Mellado Blanco, 29-52. Berlin: De Gruyter.

Jespersen, Otto. 1909-1949. *A Modern English Grammar on Historical Principles*. 7 Volumes. Copenhagen: Munksgaard. Reprint, London: George Allen & Unwin, 1961.

Poutsma, Hendrik. 1914-1929. *A Grammar of Late Modern English: For the Use of Continental, Especially Dutch, Students.* 2 Parts. 7 Volumes. Groningen: P. Noordhoff.

Semi-stable systems in PDE: paradigmatic enrichment of constructionalparadigms

Lieselotte Brems (University of Liège, Research Fellow KU Leuven)
lbrems@ulg.ac.be

This paper focuses on an underresearched but fundamental aspect of grammaticalization, namely what Diewald & Smirnova (2010) call the paradigmatic phase. Studying this paradigmatic phase is crucial to understanding how grammar and grammaticalization work; yet, it has been grossly overlooked in grammaticalization research even though it quite recently seems to be on the linguistic agenda again.

On the basis of extensive corpus extractions from the Collins Wordbanksonline corpus, this paper will look at two case studies. On the one hand, size nouns such as bunch of, lot of and load of (1-2), and on the other hand complex subordinators such as in the hope and for fear (3-4), originating in NP of NP and preposition NP complementizer syntagms respectively.

(1) People come and see a whole bunch of work, and it looks as though I work quickly (WO-UKmags)
(2) I still have a load of friends there (WO-OZnews)
(3) Where are you mum, we love you. KYLIE McDowall shudders when the phone rings for fear it could be bad news about her missing mother. (WO-OZnews)
(4) In Silicon Valley, long work hours are a badge of honor. The warp speed of innovation keeps many at full throttle for fear someone else will beat them to the next thing (WO_USNews)

I will go into the notion of 'completedness' in grammaticalization processes to show that within certain paradigms semi-stable subparadigms have existed for centuries with members in them that I will argue have grammaticalized, despite still having lexical uses and not showing full decategorialization (Hopper 1991), i.e. they are accompanied by an indefinite determiner and allow, restricted, premodification, as illustrated by (1). Rather than seeing this as incompleteness, I will argue that what could be called partial decategorialization is in fact paradigmatic enrichment (see Brems & Davidse 2010).

I will look at paradigms as constructional networks with different levels of schematicity, micro, meso- and macro level. For size nouns, the macro-level concerns the general function of quantification with regard to which size noun expressions are a meso-construction built on NP of NP syntagms. Each size noun counts as a micro-construction. Complex subordinators are subparadigms, or meso-constructions, within the paradigm of subordinators, with specific complex subordinators again functioning as micro-constructions, each displaying their own behaviour, collocational preferences and degrees of paradigmatic enrichment.

With these case studies, I zoom in on what happens in and 'after' grammaticalization, as expressions settle into a grammatical paradigm. How do specific paradigms' internal dynamics work? How are relations between potentially competing members of one paradigm (re)defined and how does a division of labour come about? I will argue that in the case studies at hand, within existing paradigms, periphrastic subsystems are integrated that are productive and semi-stable systems.

References

Collins Wordbanksonline: https://wordbanks.harpercollins.co.uk/

Brems, Lieselotte & Kristin Davidse. 2010. Complex subordinators derived from noun complement clauses.

Diewald, Gabriele & Elena Smirnova. 2010. Paradigmaticity and obligatoriness of grammatical categories. Acta Linguistica Hafniensia 42 (1): 1-10. Acta Linguistica Hafniensia 42 (2): 101-116.

Hopper, Paul J. 1991. On Some Principles of Grammaticization. In Elizabeth C. Traugott & Bernd Heine (eds.) Approaches to Grammaticalization, Volume 1. Amsterdam: John Benjamins. 17-36.

Visualizing English language: synchronic and diachronic trends

Vaclav Brezina & Raffaella Bottini (Lancaster University)
raffaellabottini@gmail.com

Visual representation of statistical information has been used extensively in many disciplines with differences in the preferred 'to go' visualization techniques, reflecting discipline-specific needs and research traditions (Tufte, 2001, 2006). This paper provides a historically grounded analysis of visualization techniques across social and natural sciences, demonstrating with examples a variety of applications in corpus linguistics with the particular focus on visual analysis of synchronic and diachronic trends in the English language. We review the historical development of visualization techniques from simple visual representations of numbers and categorisations, to displays of variables and distributions, and visualizations of complex multivariate relationships (Wainer & Velleman, 2001). Among the top ten techniques popular across social and natural sciences, we found bipartite graphs for network analyses, boxplots and scatterplots to explore variables, funnel and forest plots for meta-analyses, flow charts and argument graphs to describe different processes. In a series of case studies, we employ Brown family corpora of British English (BLOB, LOB, FLOB, BE06 and BE16) and visualize a range of synchronic and diachronic linguistic features drawing on the techniques identified in our review. In particular, we analyse and visualize the following four linguistic features: 1) modals, 2) forms of address (Mr, Mrs, Ms, Dr etc.), 3) contractions and 4) punctuation marks.

References

Tufte, E. R. (2001). The Visual Display of Quantitative Information. Graphics Press.
Tufte, E. R. (2006). Beautiful Evidence. Graphics Press.
Wainer, H., & Velleman, P. F. (2001). Statistical graphics: Mapping the pathways of science. Annual Review of Psychology, 52(1), 305–335.

Dementia metaphors in the British press: A corpus-based study

Gavin Brookes (Lancaster University)

This paper examines the most frequent metaphors that are used to represent dementia in British tabloid newspapers over a ten-year period (2010-2019). The analysis takes a corpus-based approach to metaphor identification and analysis, utilising in particular the corpus linguistic technique of collocation analysis. Metaphors are considered in terms of the 'targets' they frame, which include the following aspects of dementia: prevalence; causes; symptoms and prognosis; lived experience; responses. A range of metaphors are identified, with the tabloids exhibiting a particular preference for metaphors which construct dementia as an agentive and violent entity, people with dementia as passive victims, and which foreground preventative responses to dementia such as pharmacological intervention and individual behaviour change. It is argued that such metaphors have the potential to contribute to dementia stigma and place focus on preventing or eliminating dementia while backgrounding responses which may help people to 'live well' with the syndrome in the here-and-now. Metaphors which frame dementia as a companion or the experience of dementia as a journey are put forward as potentially less-stigmatising alternatives which might better reflect the particularities of this complex public health issue.

...because the Law commands it. A sociolinguistic study of causal conjunctions in the Old Bailey Corpus

Thomas Brunner (Catholic University of Eichstaett)
thomas.brunner@ku.de

The inventory of English causal conjunctions has been subject to striking historical changes. For instance, because, a French-derived variant introduced in Middle English, made considerable inroads in Early Modern English, and has replaced older for (that) in causal clauses and nowadays clearly outnumbers other forms such as as, since and in that (cf. Rissanen 1989; Lenker 2010; Molencki 2012). Such changes have been subject to a wide range of sociolinguistic factors. For instance, earlier research has shown genre to play a key role for the choice of causal conjunctions, with Early Modern English Bible texts favouring because more strongly than philosophical writing or trial texts, which rely on conservative for (cf. Rissanen 1998: 398; Claridge and Walker 2001: 37).

This paper sheds new light on this alternation by studying the use of causal conjunctions with regard to the influence of gender and social class in the 24-million word Old Bailey Corpus 2.0, which contains transcripts of trials from London's Central Criminal Court from 1720–1913 (Huber, Nissel and Puga 2016). On the basis of a total of 1056 conjunctions, and using multinomial regression as a method, it focuses on the following research questions:

(a) Do male and female speakers differ in their use of the causal conjunctions because, for, as, since and in that (Rissanen 1998: 398) in the course of time? Is there an evidence for women taking the lead in incoming forms (Labov 1990), especially with regard to the choice between because and for (that)?
(b) Do higher and lower social classes according to HISCO (Leeuwen & Maas 2011) differ with regard to their use of causal conjunctions, especially the introduction of because?
(c) In what way do sex and class interact?

The study expands on previous research, for one thing, by tapping a corpus which is arguably „as near as we can get to the spoken word of the period“ and by analysing data leading up to the 20th century, while previous studies have focussed on the phase until 1750 (e.g. Claridge and Walker 2001). It turns out that higher classes prefer because and later texts prefer both because and as, while gender does not play a role and, against the expection, there are no interactions.

References

Claridge, Claudia, and Terry Walker. 2001. Causal clauses in written and speech-related genres in Early Modern English. ICAME Journal 25. 31–64.

Huber, Magnus and Magnus Nissel and Karin Puga (2016). Old Bailey Corpus 2.0. hdl:11858/00-.246C-0000-0023-8CFB-2.http://fedora.clarin-d.uni saarland.de/oldbailey/downloads/OBC_2.0_Manual%202016-07-13.pdf. (25 Januar 2022).

Leeuwen, Marco H.D. van and Ineke Maas. 2011. HISCLASS: A historical international social class scheme. Leuven: Leuven University Press.

Rissanen, M. 1998. Towards an integrated view of the development of English: Notes on causal linking. In Jacek Fisiak and Marcin Krygier (eds.), Advances in English Historical Linguistics. Berlin and Boston: Mouton de Gruyter. 389–406.
Labov, William. 1990. The intersection of sex and social class in the course of linguistic change. Language Variation and Change 2. 205–54.
Lenker, Ursula. 2010. Argument and rhetoric. Adverbial connectors in the history of English. Berlin and New York: Mouton de Gruyter.
Molencki, Rafał. 2012. Causal conjunctions in mediaeval english. A corpus-based study of grammaticalization. Katowice: Uniwersytet Śląski/Oficyna Wydawnicza.

How 'real' is the quantitative turn? Investigating statistics as the 'new normal' inCorpus Linguistics

Sarah Buschfeld (TU Dortmund University), sarah.buschfeld@tu-dortmund.de
Sven Leuckert (Technische Universität Dresden)
Andreas Weilinghoff (TU Dortmund University)
Claus Weihs (TU Dortmund University), claus.weihs@tu-dortmund.de

Statistical approaches in linguistics have gained in importance in recent times, especially in the field of Corpus Linguistics. In particular, the last ten years have seen an upsurge of linguists being dedicated to statistical methods and the improvement of statistical knowledge. It seems clear that the ICAME community has long been at the front line of this statistical turn, as is apparent at its annual conferences and regular journal issues. But how 'real' is this statistical turn in quantitative linguistics (see Kortmann 2021 for a similar question)? How do statistical approaches in corpus linguistic journals differ from those journals covering related linguistic (sub)disciplines?

The present paper sets out to statistically investigate these research questions and quantitatively measure the impact of statistics in modern linguistics. To this end, we analyze the contributions to six high-impact journals regarding their use of statistical methods. The analysis covers all issues and articles published in the following journals between January 2011 and December 2021: Corpora, Corpus Linguistics and Linguistic Theory, ICAME Journal, English World-Wide, Journal of English Linguistics, and Language Variation and Change. The selected journals thematically facilitate and attract quantitative studies. At the same time, each journal comes with different linguistic breadth and foci. We built a corpus of 837 linguistic articles from these journals, MSiCoLA (Meta Studies in Corpora of Linguistic Articles), and manually browsed through all articles in search for the following criteria: Does the study employ a statistical approach? If yes, is it descriptive or inferential? Is significance considered? How is the model or approach evaluated? Is prediction considered as part of the analysis? We further noted down the exact methods and approaches employed in the studies under observation to rank them according to their frequency of use in linguistics. We model the findings for the influence of time, (Has the use of inferential statistics, model evaluation, and prediction increased in the last years?), journal, and linguistic topic (e.g. sociolinguistics, World Englishes, language acquisition, language change) by means of different inferential statistical approaches, e.g. time-series analysis, significance tests, conditional inference trees, and random forests.

Our results suggest that, indeed, inferential statistical approaches have gained ground between 2011 and 2021. They are particularly prominent in corpus linguistic journals and also depend on the linguistic topic of the respective studies. At the same time, advanced statistical approaches that employ multiple methods and include model evaluation and prediction are still rare. However, they find occasional consideration in corpus linguistics. We conclude the paper by pointing out some of these top-notch approaches, and discuss future avenues and options for advanced, statistically informed quantitative research in linguistics.

References

Kortmann, Bernd. 2021. Reflecting on the quantitative turn in linguistics. Linguistics 59(5): 1207-1226.

Diachronic Analysis of Grammatical Forms and Functions in a Corpus of 16th- to 19th-Century English Grammar Books

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

The contents of contemporary English grammar books, such as *A Comprehensive Grammar of the English Language* (Quirk et al. 1985), the *Longman Grammar of Spoken and Written English* (Biber et al. 2002), or the more recently published *Doing English Grammar* (Berry 2021) are usually organized according to parts of speech and structural elements such as phrases or clauses. Various sub-headings typically further outline, for example, different types of word class (e.g., 4. Pronouns; 4.1. Personal Pronouns; 4.2. Reflexive Pronouns). On the one hand, this modern structuring principle enhances cohesive orientation. On the other hand, the structural outline is also in line with the theoretical approach taken – functional, corpus-based etc.

This paper is a pilot study analyzing if and how Early Modern English grammarians signposted the content of their grammars through headings as cohesive devices which tie text segments together (Halliday and Hasan 1976; Fakeuade and Sharndama 2012) to "create unity of meaning" (Jambak and Gurning 2014: 61). For this purpose, a sub-corpus of these headings which will be part of the HeidelGram corpus – a representative compilation of English grammar books from the 16th until the 19th century (see e.g., Busse et al. 2020) – is compiled. Due to irregularities in typesetting, the extraction is a two-step process which relies on quantitative and qualitative methods. First, a sample of visible sign-posters in the form of section headings, which indicate to the reader what the subsequent section will be about, are identified, extracted, and quantitatively evaluated. Based on this evaluation, a larger sample is extracted in a second step for further analysis. Other types of extratextual elements such as boilerplates and notes in margins are not considered. Intratextual cohesive markers, such as topic sentences, and historiated initials are also excluded.

Based on this sample data, a diachronic analysis of the terminology used to describe grammatical categories and phenomena is performed using standard corpus linguistic tools such as WordHoard (2004-2020) which is used to track changes and salience of word-forms over time, and WMatrix (Rayson 2009) to determine key references to grammatical categories. Using modern grammatical terminology from the most commonly consulted books on English grammar (i.e., Quirk et al. 1985, Biber et al. 2002) as a baseline, we shall describe the lexico-grammatical strategies of signposting in Early Modern English grammars, thus reconstructing the development of fields of study such as morphology or syntax, and study genre conventions of English grammars in long-term diachrony. Based on this dataset of forms and functions of headings in this particular genre, we determine what grammatical categories and phenomena were most salient from the grammarians' perspective at the time, and how their centrality and representation changed diachronically.

Ultimately, this pilot study will help us in operationalizing grammatical terminology throughout time. In a follow-up study, the full grammar texts will be analyzed for their references to grammatical categories and phenomena, which will further expand the diachronic form to function mapping.

Keeping in line with the theme of the conference of whether corpus linguistics is a new normal, we portray how corpus linguistic tools enable us to efficiently and rapidly look for forms and functions in historical texts over long periods of time rather than time-consuming manual close reading.

References

Berry, Roger. 2021. Doing English Grammar: Theory, Description and Practice. Cambridge: Cambridge University Press.

Biber, Douglas, Susan Conrad, and Geoffrey Leech. 2002. Longman Student Grammar of Spoken and Written English. Harlow: Pearson Education.

Busse, Beatrix, Kirsten Gather, and 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, edited by Anna Cermakova and Markéta Malá. Diskursmuster - Discourse Patterns 20. Berlin: De Gruyter.

Fakeuade, Gbenga and Emmanuel C. Sharndama. 2012. "A Comparative Analysis of Variations in Cohesive Devices in Professional and Popularized Legal Text." British Journal of Arts and Social Sciences 4(2): 300-318.

Halliday, Michael A. K., and Ruqaiya Hassan. 1976. Cohesion in English. London and New York: Longman.

Jambak, Vany T., and Busmin Gurning. 2014. "Cohesive Devices Used in the Headline News of the Jakarta Post." Linguistica 3(1): 58-71.

Rayson, Paul. 2009. Wmatrix: a web-based corpus processing environment, Computing Department, Lancaster University. Available at http://ucrel.lancs.ac.uk/wmatrix/.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. A Comprehensive Grammar of the English Language. Harlow: Longman.

WordHoard. 2004–2020. WordHoard: An Application for the close reading and scholarly analysis of deeply tagged text. Available at http://wordhoard.northwestern.edu/userman/index.html.

The Oxford Comma in the History of English

Javier Calle-Martín (University of Málaga) & Miriam Criado-Peña (University of Granada)
jcalle@uma.es, mcriadop@ugr.es

Punctuation in historical documents has been traditionally disregarded in the literature on account of its suggested arbitrariness and the lack of correspondence to the modern system of punctuation. The Renaissance stands out as the transitional period towards the consolidation of the English system, with the establishment of the printing press contributing to some extent to the standardisation of both the inventory of symbols and the functions attributed to them. The study of historical punctuation has been mainly concerned with Old and Middle English. Even though the early Modern English has also been the object of editorial attention, most of the studies focus on literary compositions, while the other text types have been notably disregarded, scientific and legal texts in particular (Alonso-Almeida & Ortega-Barrera, 2014; Calle-Martín & Esteban-Segura, 2018). The unexplored condition of punctuation is even more significant in the particular case of early Modern printed texts, despite their active participation in the process of standardisation. Curiously enough, no studies have focused on the use of the Oxford comma in the history of English. The Oxford comma (also known as serial comma) refers to the existence of a pause immediately before the conjunctions and/or (and sometimes nor) in a series of three or more elements in a clause. Although its use is no longer a desideratum in Present-day British English, it was a disseminated practice among 17th, 18th and 19th century writers.

The present paper therefore traces the historical development of this mark of punctuation in the history of English until its eventual decline in the course of the 20th century from a corpus-based approach. In light of this, this work has been conceived with the following objectives: a) to study the use and distribution of the Oxford comma in the period 1500-1999; b) to evaluate its distribution in the two types of writing, i.e. handwriting and printing, and across text types; and c) to ascertain whether the number of elements in the series participates in its deployment. The source of evidence comes from The Málaga Corpus of Early English Scientific Prose (MCEESP), the corpus of Early English Medical Writing (CEEM) and A Representative Corpus of Historical English Registers (ARCHER 3.2.). The findings point to the impetus of the Oxford comma in the history of English as the result of eclectic forces joining their efforts at different times. The first step was taken by the early Modern English printers, who promoted its use in the 17th century, the second impulse was by the 18th- and 19th-century prescriptive grammarians, and the final step was probably taken by The Oxford English Dictionary or The Encyclopedia Britannica, which indirectly contributed to the spread of this practice throughout the 19th and the early 20th century until its eventual decline towards the middle of that same century.

References

Alonso-Almeida, F., & Ortega-Barrera, I. (2014). Sixteenth Century Punctuation in the 'Booke of Soueraigne Medicines'. Onomázein, 30, 146–168.

Calle-Martín, J., & Esteban-Segura, L. (2018). 'The Egiptians adored the Sun, and called it the visible sone of the invisible God': Clausal Boundaries in Early Modern English Scientific Handwritten Texts. Studia Neophilologica, 90(1), 68–87.

Research Trends in Corpus Linguistics: A Bibliometric Analysis of Two Decades of Scopus- indexed Corpus Linguistics Research in Arts and Humanities Introduction

Peter Crosthwaite (University of Queensland), p.cros@uq.edu.au
Sulistya Ningrum (Indonesia/State Polytechnic of Sriwijaya), arum.effendi@gmail.com
Martin Schweinberger

This paper uses a bibliometric analysis to map the field of Corpus Linguistics (CL) research in Scopus-indexed arts and humanities journal articles over the last 20 years, tracking changes in popular CL research topics, outlets, influential authors, and geographical origins of CL research.

Research questions

1. What are the most highly explored research topics in Scopus-indexed corpus linguistics research in arts and humanities between 2001-2020?
2. Who are the most contributing, most cited authors, and highest cited works by these authors during this period?
3. Which journals and countries are the most productive and influential in CL research?
4. What have been the most significant changes over the past twenty years?

Approach and Method

Bibliometric analysis identifies and tracks the long-term evolution of a thematic field, providing researchers with the key topics of the field while determining potential gaps (Groos & Pritchard, 1969: 348). Our study aimed to produce a (near) 100% sample of research developments in CL across the last 20 years in Scopus-indexed arts and humanities journal articles. Our study considers co-occurrence networks of keywords featured in CL-related keyword lists, the top-ten languages explored in CL-based research, data on the biggest contributors to the CL field by affiliation geography, and the citation rankings of the most highly cited CL researchers. We compare these metrics across four distinct time periods (2001-2005, 2006-2010, 2011-2015 and 2016-2020), charting the recent diachronic development of the field, while addressing Park and Nam's (2017: 452) call for bibliometric studies to "illustrate the corpus linguistics trends and research network among the co-cited authors", providing a birds-eye rather than a narrow view of the interdisciplinary diversity of CL research. Visualisations are produced with VOSViewer (van Eck & Waltman, 2010) with regression models performed in R.

Data

We downloaded Scopus metadata of all published journal articles in the arts and humanities subsection between 2001-2020 containing the terms 'corpus' or 'corpora' in either the title, keywords or abstract. This search included all hyphenated terms including corpus-based, corpus- assisted, corpus-derived, etc., as well as all instances of 'corpus linguistics'. Following data cleaning, exclusion of articles according to set criteria (e.g. articles without an abstract, articles where 'corpus' did not refer to a corpus of electronic language data), and an evaluation of the representativeness of our dataset against all papers published in the four main CL journals (*IJCL*, *Corpora*, *CL&LT* and *ILJCR*), information on our final dataset of **5,829** research articles spanning **97** countries, **6,379** unique authors, **193** individual language varieties, **425** individual journal outlets and **14,569** unique topic keywords are presented below:

Table 1. Dataset Information **Year**	**Search Results**	**Non CL Research**	**Abstracts Unavailable**	**Keywords Unavailable**	**Sample Dataset**
2001-2005	1,255	717	43	247	497
2006-2010	2,215	1,217	52	326	953
2011-2015	4,444	2,648	60	366	1,736
2016-2020	6,625	3,949	51	319	2,643
2001-2020	**14,539**	**8,531**	**206**	**1,258**	**5,829**

Table 2. Research Articles' Metadata **Year**	**Keywords**	**Unique Keywords**	**Languages**	**Authors**	**Journals**	**Countries**
2001-2005	1,503	908	63	613	104	47
2006-2010	4,104	2,409	85	1,223	174	58
2011-2015	7,838	4,447	93	2,297	267	76
2016-2020	13,070	6,805	112	3,528	358	90
2001-2020	**26,515**	**14,569**	**193**	**6,379**	**425**	**97**

Results

Our results reveal an increase in corpus-assisted discourse studies, lexical bundles and academic writing, a reduction (in relative terms) of studies on grammar and translation, and the introduction of new topics including multilingualism and social media. CL studies span 193 total languages/dialects across the period, and we report a significant rise in CL studies in Chinese, Russian, Spanish, and Italian over the past decade. A number of influential CL researchers have remained highly productive over the past two decades, while clusters of CL researchers are identified spanning a range of (inter)disciplinary research areas. Although the USA and the UK still account for the highest raw frequency of Scopus-indexed CL research, their slower relative increase compared with that of China, Poland, South Korea, Japan and many others is evidence that the global reach of CL research has expanded considerably over this 20-year period. Our data reveal links between developments in CL research and diachronic socio-cultural developments in applied linguistics, and society more generally. We discuss the implications of these findings for the field and provide insights into what CL research might come next.

References

Groos, O. V., & Pritchard, A. (1969). Documentation notes. *Journal of Documentation*, *25*(4), 344–349. https://doi.org/10.1108/eb026482

Park, H., & Nam, D. (2017). Corpus linguistics research trends from 1997 to 2016: A co-citation analysis. *Linguistic Research*, *34*(3), 427–457. https://doi.org/10.17250/khisli.34.3.201712.008

van Eck, N., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*, 523–538. https://doi.org/10.1007/s11192-009-0146-3

Corpus linguistics meets the law: Can an American president only be impeached for criminal conduct?

Clark D. Cunningham (Georgia State University College of Law), cdcunningham@gsu.edu
Ute Römer (Georgia State University), uroemer@gsu.edu

The U.S. Constitution states: "The President …shall be removed from Office on Impeachment for, and Conviction of, Treason, Bribery, or other high Crimes and Misdemeanors." During the first Senate impeachment trial of President Donald Trump in January 2020, emeritus Harvard law professor Alan Dershowitz argued on behalf of Trump that the words "crimes" and misdemeanors" in the Constitution are "mere synonymous terms" and therefore the charges against Trump should be dismissed without a Senate trial because the alleged conduct -- abuse of power and obstruction of Congress -- were not crimes.

The position argued by Dershowitz was widely attacked by other scholars of constitutional law, but not by arguing that "misdemeanors" functions in the text to expand the scope of impeachable offenses to non-criminal misconduct. Instead, they asserted that "high crimes and misdemeanors" was an idiomatic legal term of art, the meaning of which is determined on a case-by-case basis.

This controversy, likely to recur in future cases of impeachment, prompted us to explore the research question: "What counted as a misdemeanor in founding-era American English?" We developed a new corpus of more than 180,000 texts (over 67 million words) downloaded from the Founders Online database (founders.archives.gov) containing papers of early American leaders including George Washington, Alexander Hamilton, Benjamin Franklin, and others, written between 1706 and 1836, to explore the distributional and contextual patterns of "misdemeanor(s)" and related terms. An application of standard corpus analytic techniques including concordance, collocation, n-gram, and dispersion analysis indicated that, during the era when the US Constitution was drafted and ratified, "high crimes and misdemeanors" was not used as a fixed phrase with an idiomatic meaning. Our corpus analysis further found that "crime" and "misdemeanor" were not used synonymously in founding-era American English. Having found evidence that counters both of the competing interpretations advanced by American legal scholars, we then asked how "other" and "high" in the impeachment clause function as modifiers and concluded that both likely modified "misdemeanors" as well as "crimes."

If impeachment can be based on either "high crimes" or "high misdemeanors," the course of future impeachment cases could turn on the meaning of "high misdemeanors." Our research as of the date of this proposal indicates that "high misdemeanor" appears to be a regular phrase in founding-era American English, and is not simply derived from the impeachment clause. The phrase most frequently appears in contexts describing misconduct that affects government functioning or threatens the authority or security of the state and is sometimes used to describe misconduct that justifies removal from office. However, instances in the Founders Online database are limited, so we are currently expanding our research to additional corpora that are contemporaneous with or predate the ratification of the Constitution, including COEME (Corpus of Early Modern English), COFEA (Corpus of Founding Era American English), and subsets of COHA (Corpus of Historical American English).

Combining corpus and qualitative methods to improve the representation of spoken language in ELT materials

Niall Curry (Coventry University)
ad3377@coventry.ac.uk

In published ELT classroom materials, norms relating to written language often dominate at the expense of spoken language. However, there is a perennial request from learners for more 'conversation' in materials, owing to general feelings of anxiety surrounding spoken performance. At the core, the features characteristic of everyday spoken language do not reflect the rather neat syllabi expected by teachers and students, globally. Moreover, there are deeply-held opinions relating to usage and acceptability of such features of spoken language, which are often seen as ungrammatical or overly complex. As a result, features of spoken conversation are often missing from mainstream materials. Overall, the challenge is to include a focus on spoken language in ELT materials that (1) does not require much space on a page, (2) can fit within teacher expectations, (3) can help learners improve their conversational competencies, and (4) does not undermine learner success in language assessments. To address this challenge, our research combines a corpus-based conversation analysis of 'small words' (Carter & McCarthy, 2017) with qualitative workshops and focus groups with teachers, editors, and assessment developers to gain a comprehensive perspective from core stakeholders in ELT materials development.

To address the aims of this project, the proposed study is guided by the following research question:

1. What are the core features of casual spoken conversation, how do they feature in ELT materials, and to what degree is corpus linguistics seen to be a valuable resource for improving this representation?

This project is composed of three distinct phases: 1) corpus analysis, 2) coursebook/materials review, and 3) workshops with ELT practitioners. To conduct the analysis, first, the corpus research involved analysing spoken corpora (e.g. Cambridge Reference Corpus) to study turn-boundaries and language patterns to identify a list of features of spoken conversation. Second, for the evaluation of materials, we drew on the findings of the corpus analysis to code ELT materials using critical grounded theory (Hadley, 2017). We then extracted key insights regarding the presence or lack thereof of identified features of spoken language in the ELT materials to develop workshops. The workshops asked participants to design spoken lessons and consider the value of corpus linguistics for materials development, for example. In analysing the workshop data, we code the both the transcripts of discussions and the lesson materials produced.

A number of key findings have emerged from the analysis to-date. These include corpus findings indicating that yes/no questions are rarely answered with a yes or no – something not always seen the ELT materials. Moreover, practitioner reflections on the value of frequency information, the role of small words in generating challenge, and the value of corpus linguistics for training and development emerged. Overall, this project brings together a set of perspectives on ELT materials that, heretofore, have never been studied together. Furthermore, the focus on corpus linguistics offers a valuable opportunity to address the shortcomings of the ongoing corpus revolution (Rundell &

Stock, 1992; Chambers, 2019) and to develop strategies for exploiting corpora better for the future of ELT materials development.

References

Carter, R., & McCarthy, M. (2017). Spoken grammar: Where are we and where are we going?. Applied linguistics, 38(1), 1-20.

Chambers, A. (2019). Towards the corpus revolution? Bridging the research–practice gap. Language Teaching, 52(4), 460-475.

Hadley, G. (2017). Grounded theory in applied linguistics research: A practical guide. Routledge.

Rundell, M., & Stock, P. (1992). The corpus revolution. English Today, 8(3), 21-32.

Linguistic factors in successful persuasion online

Daria Dayter (Tampere University)

The present paper is a part of a larger project that aims to answer the following research questions: What linguistic features characterize successful persuasion in social media environments? and, more specifically, Can linguistic variation be detected in a contrasting subcorpora or successful and unsuccessful persuasive Reddit posts?

The source of the data for the study is the subreddit R/ChangeMyView (CMV), a community of users revolving around the practice of discursive persuasion. The 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.

This paper builds on earlier research (Dayter and Messerli 2021a, 2021b) investigating persuasive practices on CMV. As that study demonstrated, no meaningful linguistic variation can be observed when comparing the DAC and non-DAC subcorpora: both successful and unsuccessful persuaders use formal, standard English, appeal to external evidence, and exhibit a high rate of linguistic features associated with Biber's Overt Persuasion dimension.

The paper takes the project to its next step, contrasting DAC and non-DAC comments in subcorpora organised per thread. Such a setup allows the researcher to trace the persuasive language and persuasion strategies that succeeded (or failed) with the particular post author, thus taking out idiosyncratic malleability of opinion as a confounding factor.

The resulting dataset comprises 100 initial posts and the corresponding threads selected from the larger CMV corpus, with the average length of 6,000 tokens per thread. Keyword and N-gram analysis of the DAC and non-DAC comments served as the starting point, with the positive keywords grouped into topoi and concordances for each topos manually examined and annotated. A subsample of 10 threads had undergone qualitative analysis based on existing taxonomies of persuasive strategies (Guadagno and Cialdini 2009, Locher 2006).

The findings confirm that per-thread design is the more adequate choice for the contrastive study of persuasion on Reddit. The list of keywords extracted based on LL and %DIFF measures yielded a split in malleability of opinion along two broad thematic groups. The DACs can be grouped under a broad label of ‚entertainment' (opinions about film, books, popular culture, space exploration). The non-DAC can be grouped under the ‚politics and religion' label. Similar split can be traced when comparing the N-Gram lists. Finally, the qualitative analysis of persuasive strategies showed a slight preference towards constructing identity of an expert using external evidence in politically-themed threads.

References

Dayter, Daria and Thomas C. Messerli. 2021a. Online First. "Persuasive language and features of formality on the r/ChangeMyView subreddit." Internet Pragmatics https://doi.org/10.1075/ip.00072.day

Dayter, Daria and Thomas C. Messerli. 2021b."Persuasive register in the Change My View subreddit" Presentation at the iCAME42 conference, Dortmund University, Germany, August 18-21, 2021.

Guadagno, R. & R. Cialdini. 2009. "Online Persuasion and Compliance: Social Influence on the Internet and beyond". In The social net: The social psychology of the Internet, ed. by Y. Amichai-Hamburger. New York: Oxford University Press.

Locher, M. 2006. Advice Online. Amsterdam: Benjamins.

The crisis of normality. Conceptual metaphorical patterns in the discourse of crisis: new old mappings

Dario Del Fante (University of Padova)

Recently, different crises have globally posed challenges to the stability of our contemporary societies: Covid-19 sanitary crisis, 2015 migration crisis and the climate crisis. The term “crisis” comes from the Latinized form of Greek “κρίσις” (krisis). It refers to a turning point in a disease or to that change that indicates recovery or death. Specifically, it derives from the verb “κρίνω” (krino) which means "to separate, decide". Following this, a crisis can be interpreted as a separated moment from a period of stability. In this sense, a moment when norms might be suspended or may be subject to variations: normality is not normal anymore and an ab-normality has to be established (until it becomes usual and then normal). Particularly, when faced with crises such as sanitary or social ones, metaphors and commonplace images are often used to conceptualize and communicate about them in media and political discourse (Charteris-Black 2021).

Metaphors play a fundamental role in understanding and influencing how we think and talk about reality. Human reasoning is intrinsically metaphorical and imaginative, metaphors connect the domain of concrete and distinct experiences (the Source Domain) onto the domain of predominantly abstract and complex experiences (the Target Domain), thus enabling us to better understand the complexity of reality that surrounds us (Semino 2008). A crisis is a kind of subjective experience that tends to be talked about by means of metaphorical expressions. In this sense, the analysis of conceptual metaphorical patterns used to communicate about a crisis can help us to advance our understanding of how we interact with and react to these problematic events and how we conceptualize an interruption of normality.

To address this issue, I intend to embark on a case study: drawing on previous metaphor research on migration (Charteris-Black 2006; Taylor 2021) pandemics (Semino 2021) and climate change (Shaw & Nerlich 2015; Adam & Wahyuni 2020) and building on conceptual metaphor theory (Kövecses 2020), this project uses Corpus-assisted Critical Metaphor Discourse Analysis (Charteris-Black 2004) to examine the metaphorical representation of three recent crises in newspaper and parliamentary discourse: Covid-19 sanitary crisis, climate change and 2015 migration crisis. Lastly, the paper aims to investigate newspapers and political communication within a not-normal situation.

Three datasets will be analysed:

- A collection of articles on Migration from two UK newspapers and two Italian newspapers published between 2015 and 2016 (10 million tokens each).
- A collection of articles on climate change from two UK newspapers and two Italian newspapers published between 2019 and 2021 (10 million tokens each).
- A collection of articles on Covid-19 from two UK newspapers and two Italian newspapers published between 2020 and 2022 (10 million tokens each);

A cross-linguistics perspective has been adopted to expand the scope of our research and to let comparison among two countries that are both strongly connected to the crises under study.

Preliminary results suggest the presence of three main metaphorical mappings within both migration discourse and pandemic discourse, whilst only one is also shared with climate change discourse:

- WATER: wave /surge of/flow/rise+ migrant/Covid-19;
- FIRE: explosion of + migrant/Covid-19;
- WAR: invasion of/attack/fight/combat + migrant/ Covid-19 /Climate Change

References

Adam, M., & Wahyuni, W. (2020). The Image of Climate Crisis in Media: A Conceptual Metaphor Analysis. Journal of Language and Literature, 20(1), 10–24. https://doi.org/10.24071/joll.v20i1.2413

Charteris-Black, J. (2004). Corpus approaches to critical metaphor analysis. Basingstoke: Palgrave-MacMillan.

Charteris-Black, J. (2006). Britain as a container: Immigration metaphors in the 2005 election campaign. Discourse & Society, 17(5), 563-581.

Charteris-Black, J. (2021). Metaphor of Coronavirus. Invisible Enemy or Zombie Apocalypse? Basingstoke: Palgrave-MacMillan.

Kövecses, Z. (2020). Extended Conceptual Metaphor Theory. Cambridge: Cambridge University Press.

Semino, E. (2008). Metaphor in discourse. Cambridge: Cambridge University Press.

Semino, E. (2021): " 'Not Soldiers but Fire-Fighters' – Metaphors and Covid-19", In: Health Communication, 36, 1, 50–58.

Shaw, C., & Nerlich, B. (2015). Metaphor as a mechanism of global climate change governance: A study of international policies, 1992–2012. Ecological Economics, 109, 34–40. https://doi.org/10.1016/j.ecolecon.2014.11.001

Taylor, C. (2021): "Metaphors of Migration over Time", In: Discourse & Society, 1-19.

Reporting clauses in British general vs. crime fiction

Jarle Ebeling & Signe Oksefjell Ebeling
jarle.ebeling@gmail.com; s.o.ebeling@ilos.uio.no

The study of fictional dialogue and reported speech has a fairly long tradition in literary linguistics, most notably the study of the meaning and use of reporting verbs and clauses in 19th century fiction, and in Dickens's novels in particular (e.g. Lambert 1981, Mahlberg 2013). A recent study of reporting clauses in 20th century British fiction (Ebeling 2021) shows how the use of SAY as a reporting verb has increased at the expense of semantically fuller verbs, such as CRY and MUTTER, over the past century. At the same time the use of manner adverbs to qualify the act of reporting, e.g. "Look out, Joan," said Vane nervously., seems to have declined, while more elaborate glossing phrases (Caldas-Coulthard 1987) such as, "You people never stop," he moaned, bubbling through his left nostril., seem to be on the rise. Such accompanying circumstances to the speech act itself is the main focus of this paper.

The paper starts out with an overview of the nature of reporting clauses in 20th and 21st century British fiction overall, including the use of adverbs and different types of accompanying circumstances. Then the study narrows its scope to offer a more detailed description of the use of adverbs and non-finite -ing clauses, with a focus on the past 40 years and two literary genres, namely crime and general fiction. Thus, the study employs methods and approaches to the study literary language drawn from distant reading and corpus stylistics, as well as corpus linguistics proper through close reading of concordance lines. The data are culled from the 100-million word Corpus of British Fiction[2], consisting of more than 1,200 novels published between 1900 and 2019.

Tracking the use of reporting clauses over the last few decades will enable us to answer the following research questions: 1) to what extent does literary subgenre contribute to the makeup of such clauses?, and 2) to what extent have the two subgenres developed differently in recent years?

A preliminary investigation of 1,000 random instances of direct speech in two decades, viz. the 1980s vs. the 2010s in general vs. crime fiction shows that, while the use of single adverbs seems to be declining in general fiction their use is stable in crime fiction. The use of non-finite -ing clauses as an accompanying circumstance is shown to decline modestly in general fiction but shows a sharp rise in the crime novels. Further scrutiny of -ing clauses in the two genres will be conducted in order to enhance our understanding of how, and potentially also why, the two genres differ in this regard. Thus, we will investigate a specific linguistic feature that may set the two literary genres apart in terms of style, along the lines of advice given to writers regarding the use of adverbs: "the general rule in fiction is to eliminate as many adverbs as possible"[3], and "[n]ever use an adverb to modify the verb 'said'"[4].

References

Caldas-Coulthard, Carmen Rose. 1987. Reported speech in written narrative texts. In Malcolm Coulthard, (ed.), Discussing Discourse. Studies Presented to David Brazil

[2] https://nabu.usit.uio.no/hf/ilos/cbf/cbfhelp.html
[3] https://www.autocrit.com/editing/support/adverbs-in-dialogue/
[4] https://www.theguardian.com/books/2010/feb/20/ten-rules-for-writing-fiction-part-one

on his Retirement, Birmingham: English Language Research, University of Birmingham, 149–167.
Ebeling, Jarle. 2021. 120 years of reporting clauses: stability or change? Paper read at ICAME 21, Dortmund.
Lambert, Mark. 1981. Dickens and the Suspended Quotation. New Haven: Yale University Press.
Mahlberg, Michaela. 2013. Corpus Stylistics and Dickens's Fiction. London: Routledge.

Determining Letter-Specific Speech Acts in 18th Century Varieties of English

Christine Elsweiler & Patricia Ronan (TU Dortmund University)
patricia.ronan@tu-dortmund.de

Corpus-based approaches to pragmatics have recently seen notable advances (e.g. Rühlemann & Aijmer 2015, Kallen & Kirk 2012). However, corpus-based variational pragmatic approaches are still rare, and particularly in historical pragmatics the research base is slim so far. In particular methods which allow for structured, reproducible results are still a desiderate. The proposed paper intends to fill this gap by investigating variation in 18th century letter-specific speech acts, viz. greeting and leave-taking formulae in 18th century Scottish, English and Irish letters.

Previous studies have shown that greeting and leave-taking formulae in historical English letters underwent diachronic changes between the 15th and the late 17th century. Salutation strategies were e.g. structurally simplified from complex formulae such as *Right honourable and worshipful Sir* to simple *Sir* (Nevalainen and Raumolin-Brunberg 1995). Moreover, in the course of the early modern period, letter-writers manifested a trend moving from negative politeness strategies towards the use of positively polite greeting and leave-taking formulae evincing solidarity and affection between the interlocutors, e.g. *Your most affectionat sister* (Nevala 2003). While these studies considered sociopragmatic factors such as social distance and power, to date, regional differences regarding the choice of greeting and leave-taking formulae have not been explored. It has been shown for early modern Scottish and English letters, though, that the macro-social factor region may influence the choice of pragmalinguistic strategies (author a). Moreover, there is evidence for cross-cultural differences in the realisation of letter closings in 19th century British and German letters (House and Kádár 2021: 207–215).

The current study adds to previous research by investigating features of and differences in interpersonal interaction evidenced in greeting and leave-taking formulae in 18th century letters. Specifically, it pursues the research questions which letter-specific speech acts can be found in 18th century Scottish, English and Irish letters and how they compare in the three varieties at hand. To answer the research questions, we analyze a sample of 100 18th century letters each of typically upper and upper-middle class writers from Scotland, England and Ireland. The data set consists of a random selection of 100 private and non-private letters of the Scottish corpus component, *ScotsCorr*. To these we add a random selection of 100 similar Irish-authored 18th century letters and an equal number of English-authored letters addressed to other English recipients, which are held at the National Library of Ireland. Data are manually annotated for the letter-specific speech act patterns salutations and leave taking. The approach uses a categorization frame that determines formal and functional criteria for all observable speech acts, building on and extending Blum-Kulka et al.'s framework for speech act sequences (author a,b). It thus offers the possibility to maximize reproducibility and accountability so that the approach can also be replicated in research on other historical or contemporary varieties and thus allow for maximum comparability.

The results determine the formulae that are in use in the corpus materials. They show to what extent the greeting and leave-taking formulae differ in the varieties under investigation. The approach will facilitate further synchronic and diachronic work in corpus pragmatics.

References

Elsweiler, Christine. 2022. “Gender Variation in the Requestive Behaviour of Early Modern Scottish and English Letter-writers? A Study of Private Correspondence”. *Journal of Historical Sociolinguistics* 8.1: 55–88.

Elsweiler, Christine. Forthcoming. “Modal *May* in Requests: A Comparison of Regional Pragmatic Variation in Early Modern Scottish and English Correspondence”. *Journal of Historical Pragmatics* 26.2.

Blum-Kulka, Shoshana, Juliane House and Gabriele Kasper. 1989. “The *CCSARP* Coding Manual”. In: Shoshana Blum-Kulka, Juliane House and Gabriele Kasper (eds.). *Cross-cultural Pragmatics: Requests and Apologies*. Norwood, NJ: Ablex Publishing Corporation. 273–294.

House, Juliane and Daniel Kádár. 2021. *Cross-cultural Pragmatics*. Cambridge: Cambridge University Press.

Kallen, Jeremy and John Kirk. 2012. *SPICE-Ireland: A User’s Guide*. Belfast: Cló Ollscoil na Banríona.

Nevala, Minna. 2003. “Family First: Address and Subscription Formulae in English Family Correspondence from the Fifteenth to the Seventeenth Century”. In: Irma Taavitsainen and Andreas Jucker (eds.). *Diachronic Perspectives on Address Term Systems*. Amsterdam and Philadelphia: Benjamins. 147–176.

Nevalainen, Terttu and Helena Raumolin-Brunberg. 1995. “Constraints on Politeness: The Pragmatics of Address Formulae in Early English Correspondence“. In: Andreas Jucker (ed.). *Historical Pragmatics: Pragmatic Developments in the History of English*. Amsterdam and Philadelphia: Benjamins. 541-601.

Rühlemann, Christoph and Karin Aijmer. 2015. “Corpus Pragmatics: Laying the Foundations”. In: Karin Aijmer and Christoph Rühlemann (eds.). *Corpus Pragmatics: A Handbook*. Cambridge: Cambridge University Press. 1–26.

ScotsCorr = The Helsinki Corpus of Scottish Correspondence 1540–1750. 2017. Ed. Anneli Meurman-Solin. Helsinki: University of Helsinki. <http://urn.fi/urn:nbn:fi:lb-201411071>.

Capturing and Analysing Multimodality in a Corpus of Royal Correspondence: A Case Study of the Letters of James V and Henry VIII

Mel Evans (University of Leeds) & Helen Newsom (Aston University)
m.evans5@leeds.ac.uk; h.newsome@aston.ac.uk

Naturalization is a key component in the establishment and maintenance of hegemonic discourses and ideology (Fairclough 1985). Studies of present-day and historical texts demonstrate that the construal of 'normality' is achieved through multiple semiotic modes, including language, which accrues ideological force through its persistence over time and across genres (e.g. Taylor 2020). This paper interrogates the construction of 'normality' and naturalization in the epistolary practices of kingship in Scotland and England (1513-1542). Correspondence was a vital tool in the diplomatic negotiations between James V of Scotland and Henry VIII of England; a period of both war and peace between the two realms. As the substitute for face-to-face negotiations, letter-writers developed social practices for identity construction and interpersonal relationship, via linguistic (e.g. salutations, formulae) and material (e.g. signature placement, white space) means. The epistolary performance of kingship, therefore, was critical. Although their correspondence was in mutually intelligible varieties (Scots and English), the letters of James V and Henry VIII originated from different (often conflicting) cultural positions. It is unclear the extent to which the norms of epistolary kingship in Scotland may have been shared with those of England, and how those norms may have changed over the thirty-years-long interaction between the two adjoining realms.

The investigation of historical correspondence using corpus linguistic approaches is well established (e.g. Nevalainen and Raumolin-Brunberg 2003). However, in accordance with the 'material turn' in early modern studies (e.g. Daybell 2012) and recognising that discourses are embedded in their contexts of use (Van Dijk 2006), the language of these letters is only part of the resources used to create meaning and construe the authority of their named authors. Our research questions for this paper are therefore primarily methodological:

1) How can material, linguistic and situational (extra-textual) information be effectively captured in a corpus of historical correspondence?
2) Which methods are most appropriate for the identification of the linguistic and material construal of royal authority in the corpus?
3) What semiotic features characterise royal correspondence in the corpus, and how do these contribute to the naturalization of monarchic power within and across Scottish and English realms?

Our paper first reports on the process of corpus creation, based on the images and transcripts of 200 letters issued in the name of the Scottish and English kings. We address the challenges of access to, and the representativeness and reliability of these UK archival materials, held on-site and online (e.g. State Papers Online) for corpus analysis, and outline how our corpus integrates linguistic and material elements through a combination of TEI XML mark-up, metadata, and linked images. Finally, we discuss our preliminary findings based on keywords and n-gram analyses of the material-linguistic properties, and the challenges of interpretation. We argue that the naturalization of royal discourses hinges on the interplay between situational, material and linguistic forms. For Scottish and English kingship, these discourses show significant deviations along national lines,

with implications for the long-term trajectory of cultural and linguistic ideologies of power.

References

Daybell, J. 2012. The Material Letter in Early Modern England [Online]. London: Palgrave Macmillan UK.

Fairclough, N.L. 1985. Critical and descriptive goals in discourse analysis. Journal of Pragmatics. 9(6), pp.739–763.

Nevalainen, T. and Raumolin-Brunberg, H. 2003. Historical sociolinguistics: Language Change in Tudor and Stuart England. London: Pearson Education.

Taylor, C. 2021. Metaphors of migration over time. Discourse & Society. 32(4), pp.463–481.

Van Dijk, T.A. 2006. Ideology and discourse analysis. Journal of Political Ideologies. 11(2), pp.115–140.

A corpus-based analysis of Irish English speakers' virtual intercultural communications in the technology sector

Gail Flanagan (University of Limerick)
Gail.Flanagan@ul.ie

This study investigates the intercultural communications of Irish English speakers who work in the Irish technology sector. The technology sector employs over 210,000 staff in Ireland today with 9 out of 10 global software and US technology companies basing their European headquarters in Ireland (Technology Ireland report, 2020). The goal of this research is to identify key features in Irish English speech in international virtual teams which in turn, will scaffold the creation of *on the job* based intercultural communication training for higher education and professional learners. The study focuses on virtual teams which are already established as the norm in the technology sector with remote, rather than face-to-face, interactions, having further increased substantially with the travel restrictions due to the Covid-19 pandemic. Although noted that the research centres on the verbal pragmatic traits of Irish English interlocutors, the corpus will consist of both Irish and International participants, thereby facilitating research findings across international business teams.

While there exists substantial research around business discourse (Drew and Heritage, 1992) and virtual teams (Ford, 2017; Lockwood, 2015), many studies address the challenges associated with managing, rather than participating in, such teams. This research proposes a bottom-up approach, targeting individual contributors, as it is in this role that most Irish employees begin their working lives. This researcher also intends to solidly contribute to Business English as a Lingua Franca (BELF) pragmatics theory (building on Seidlhofer, 2004) with the expansion to include Irish English speakers and furthermore, address the paucity of Irish workplace discourse studies (a notable exception being Cacciaguidi-Fahy and Fahy, 2005).

The research methodology involves the creation of a transcribed corpus of Business English as a Lingua Franca (BELF) speech of approximately 150,000 words. The data will be transcribed from web-based recordings of international virtual meetings that include Irish English speakers. Conversation Analysis (CA) techniques will be used to qualitatively analyse the transcribed speech, further supported by a quantitative analysis using corpus linguistics methods. Furthermore, the corpus-based findings will be compared with the results from a widely distributed communication behaviour survey (completed in 2021). This comparative analysis will identify any delta between Irish English interlocutors' perception of their intercultural communication behaviour and the reality as evidenced in the *language in action* spoken corpus analysis.

References

Cacciaguidi-Fahy, S., & Fahy, M. (2005). Whatcha mean? The pragmatics of intercultural business communication in financial shared services centres. In A. Barron and K. Schneider (Eds.), *The pragmatics of Irish English* (pp. 266-312). Berlin: Mouton de Gruyter.

Drew, P., & Heritage, J. (1992). *Talk at work: Interaction in institutional settings*. Cambridge: Cambridge University Press.

Ford, R.C., Piccolo, R.F. & Ford, L.R. (2017). Strategies for building effective virtual teams: Trust is key. *Business Horizons*, *60*(1), 25-34. Retrieved from https://doi.org/10.1016/j.bushor.2016.08.009

Lockwood, J. (2015). Virtual team management: What is causing communication breakdown?. *Language and Intercultural Communication, 15*(1), 125-140. doi: 10.1080/14708477.2014.985310.

Seidlhofer, B. (2004). Research perspectives on teaching English as a lingua franca. *Annual Review of Applied Linguistics*, *24*, 209-239. Retrieved from https://doi.org/10.1017/S0267190504000145.

Technology Ireland (2020). *Future needs, future thinking 2020.* Online report. Retrieved from https://www.technology-ireland.ie/Sectors/TI/TI.nsf/vPages/About~Press~future-needs,-future-thinking-2020-04-11-2019/$file/TI+Future+Needs,+Future+Thinking.pdf

A quasi-longitudinal analysis of the L2 acquisition of tense and aspect

Robert Fuchs & Valentin Werner
robert.fuchs.dd@gmail.com

Our paper expands on previous work in the area of morphosyntax (see, e.g., Deshors, 2018, 2021; Li, 2020; Werner et al., 2021 and contributions to Ayoun, 2015; Howard & Leclercq, 2017; McManus et al., 2017; Fuchs & Werner, 2020) with a view to testing and refining established SLA principles on the acquisition of tense and aspect (TA) markers. Specifically, we consider (i) the order of acquisition of tense and aspect (OATA) and (ii) the Default Past Tense Hypothesis (DPTH). To date, these hypotheses have been put to the test only in smaller learner groups, mainly applying experimental SLA approaches (see, e.g., Bardovi-Harlig, 2000; Salaberry, 2008; Svalberg, 2018; O'Reilly, 2018; Jiránková & Cilibrasi, 2021).

Proponents of the OATA (see, e.g., Bardovi-Harlig 2000; Svalberg 2018) agree on an emergence of TA forms in learner English along the following lines: simple present/present progressive > simple past/past progressive > present perfect > present perfect progressive > past perfect > past perfect progressive. Proponents of the DPTH (e.g. Salaberry & Ayoun 2005) predict that learners in early-intermediate stages will use a single morphological marker for past-time reference, which for EFL learners is the simple past.

In this paper, we test the predictions of the OATA and DPTH on data from learners of English as a Foreign Language at school and university level. The central issue in focus is to what extent an increase in the frequency of usage corresponds to an increase in accuracy.

Accordingly, we use a quasi-longitudinal research design and measure both the frequency and the accuracy of usage of TA markers, using multi-layer error annotations to explore whether and to what extent an increase in the frequency of usage corresponds to an increase in accuracy. Data is drawn from the International Corpus of Crosslinguistic Interlanguage (Tono & Díez-Bedmar, 2014) and the International Corpus of Learner English (Granger et al., 2009) to assess TA acquisition in (tutored) learner writing from the beginning to the advanced level in four typologically different L1 backgrounds (German, Chinese, Polish, Spanish). Based on the categories established in Dagneaux et al. (2005), error ratings of more than 4,000 data points (verb tokens) were provided by two native speakers, with disagreements between these raters being resolved by a third native-speaker rater.

In the larger picture, our findings confirm the predictions of the OATA and the DPTH. Findings indicate that simple forms are used (i) earlier and more frequently and (ii) more accurately than complex forms at any stage in the acquisition process. However, the data also are also suggestive of nuanced patterns: Results indicate that accuracy of usage does not linearly increase with frequency of usage or proficiency.In addition, the manual accuracy ratings allow us to assess (i) accuracy of usage in terms of "false negatives" (e.g. using a present simple where a present progressive is required) and (ii) particular error types (functional errors – i.e. confusion of TA forms – and formal errors – e.g. omission of 3rd person singular -s in the present).

References

Ayoun, D. (2015). The acquisition of the present. Amsterdam: Benjamins.

Bardovi-Harlig, K. (2000). Tense and aspect in second language acquisition: Form, meaning, and use. Malden: Blackwell.
Dagneaux, E., Dennes, S., Granger, S., Meunier, F., Neff, J., & Thewissen, J. (2005). Error tagging manual version 1.2. Louvain: Centre for English Corpus Linguistics.
Deshors, S. C. (2018). Does the passé composé influence L2 learners' use of English past tenses? International Journal of Learner Corpus Research, 4(1), 23–53.
Deshors, S. C. (2021). Contextualizing past tenses in L2: Combined effects and interactions in the present perfect versus simple past alternation. Applied Linguistics, 42(2), 269–291.
Fuchs, R., & Werner, V. (2020). Tense and aspect in second language acquisition and learner corpus research. Amsterdam: Benjamins.
Granger, S., Dagneaux, E., Meunier, F., & Paquot, M. (2009). The International Corpus of Learner English: Version 2. Louvain-la-Neuve: Presses universitaires de Louvain.
Howard, M., & Leclercq, P. (2017). Tense-aspect-modality in a second language: Contemporary perspectives. Amsterdam: Benjamins.
Jiránková, L., & Cilibrasi, L. (2021). Second-language acquisition of the English past-tense: From rules to analogy. Linguistica Pragensia, 31(2), 188–213.
Li, J. (2020). Tense and aspect in second language acquisition: A corpus-based comparison of Chinese and German learner English. PhD thesis, University of Freiburg.
McManus, K., Vanek, N., Leclercq, P., & Roberts, L. (2017). Tense, aspect, and modality in L2. Special issue of the International Review of Applied Linguistics in Language Teaching, 55(3).
O'Reilly, J. (2018). Processing and production of unique and non-unique-to-L2 syntactic structures: The case of English articles and tense-aspect. PhD thesis, University of York.
Salaberry, R. (2008). Marking past tense in second language acquisition: A theoretical model. London: Continuum.
Svalberg, A. M.-L. (2018). Mapping tense form and meaning for L2 learning – from theory to practice. International Review of Applied Linguistics in Language Teaching, 57(4), 417–445.
Tono, Y., & Díez-Bedmar, M. B. (2014). Focus on learner writing at the beginning and intermediate stages: The ICCI corpus. International Journal of Corpus Linguistics, 19(2), 163–177.
Werner, V., Fuchs, R., & Götz, S. (2021). L1 influence vs. universal learning mechanisms: An SLA-driven corpus study on temporal expression. In B. Le Bruyn & M. Paquot (eds.), Learner corpus research meets second language acquisition. Cambridge: Cambridge University Press. 39-66.

The Dative Alternation in Pre- and Post-Handover Hong Kong: Towards Endonormative Stabilization or Restriction?

Nina Funke (Justus Liebig University Giessen)
Nina.Funke@anglistik.uni-giessen.de

Hong Kong English (HKE) has entered the nativization phase of Schneider's (2007) dynamic model of the evolution of postcolonial Englishes in the 1960s, i.e. the phase in Schneider (2007) which shares certain characteristics with the phases of expansion and institutionalization in Moag's (1982, 1992) life cycle of non-native Englishes. However, since the handover of Hong Kong to China in 1997, the Chinese government has been promoting the use of Cantonese over English (e.g. Bolton et al. 2020). As said language policies may affect its nativization process, HKE might be developing towards a phase of restriction characterised by Moag (1982) as a local reversion of English to EFL status – in contrast to the evolutionary path towards endonormative stabilization suggested by Schneider (2007). Against this background, the present study investigates the short-term diachronic development of the dative alternation, i.e. the choice between the ditransitive (1) and the prepositional (2) dative. While the dative alternation has been studied in HKE as well as other Southeast Asian Englishes (e.g. Röthlisberger 2018), these studies are synchronic analyses of this structural feature. Therefore, the present study serves as a starting point to the diachronic analysis of structural nativization of HKE based on the dative alternation.

(1) […] Mr Ma owed citizens an apology […] (SCMPC_2002-10-12_58)
(2) […] people give the credit to the guys […] (SCMPC_1993-01-16_163)

The analysis is based on newspaper language sampled from the South China Morning Post (SCMP) before and after the handover – from 1993 and from 2002. The dative alternation is studied in relation to these research questions:

1. Did the use of datives in HKE change from 1993 to 2002; and if so, did it move towards endonormative stabilization (cf. Schneider 2007) or restriction (cf. Moag 1982)?
2. What factors cause differences in the dative use between HKE and British English (BrE) speakers?

2400 interchangeable datives were extracted from the SCMP Corpus as compiled at Justus Liebig University Giessen and from the British National Corpus (BNC). Each dative was annotated for twelve different structural, e.g. constituent length, and context-related, e.g. lexical density, variables. A Multifactorial Prediction and Deviation Analysis with Random Forest (MuPDARF; Deshors & Gries 2016) identifies LENGTH, FREQUENCY, and SURPRISAL as significant predictors of diachronic change in the HKE use of datives. While the importance of constituent LENGTH was established by previous research, SURPRISAL has not been systematically considered in studies of the dative alternation. It also emerges that the data taken from the BNC is able to predict the HKE dative choice in 2002 significantly better than in 1993. This indicates that instead of institutionalising localised quantitative preferences as characteristic of an endonormatively stabilized variety following Schneider's (2007) model, HKE is moving towards the phase of restriction described in Moag's (1992: 245) life cycle. The fact that HKE dative use in 2002 is more similar to BrE in 2002 than it was in 1993 shows a stronger exonormative influence from

BrE and in the long run possibly a turn towards an EFL status as described by Moag (1992).

References

Bolton, Kingsley, John Bacon-Shone, and Kang Kwong Luke (2020). "Hong Kong English." The Handbook of Asian Englishes. Eds. Kingsley Bolton, Werner Botha, and Andy Kirkpatrick. Hoboken: John Wiley & Son. 449—478.

Deshors, Sandra C., and Stefan Th. Gries (2016). "Profiling Verb Complementation Constructions across New Englishes: A Two-Step Random Forest Analysis of ing vs. to Complements." International Journal of Corpus Linguistics 21(2): 192—218.

Levy, Roger (2008). "Expectation-based Syntactic Comprehension." Cognition 106(3): 1126—1177.

Moag, Rodney F. (1982). "The Life Cycle of Non-native Englishes: A Case Study." The Other Tongue: English across Cultures. Ed. Braj B. Kachru. Oxford: Pergamon Press. 270—288.

Moag, Rodney F. (1992). "The Life Cycle of Non-native Englishes: A Case Study." The Other Tongue: English across Cultures, 2nd edition. Ed. Braj B. Kachru. Urbana IL: University of Illinois Press. 233—252.

Röthlisberger, Melanie (2018). Regional Variation in Probabilistic Grammars: A Multifactorial Study of the English Dative Alternation. Doctoral Dissertation. Leuven: KU Leuven.

Schneider, Edgar W. (2007). Postcolonial English: Varieties around the World. Cambridge: CUP.

Triangulating methods in corpus linguistics: from frequency to move and dialogic analysis

Matteo Fuoli (University of Birmingham) & Monika Bednarek
m.fuoli@bham.ac.uk

Recent years have seen growing recognition of the benefits of methodological triangulation in corpus-linguistic research, whether it involves the combination of multiple corpus analysis techniques (e.g. Baker & Egbert 2016) or the integration of corpus analysis with other research methods, such as qualitative interviews (e.g. Bednarek 2019) or experiments (e.g. Fuoli et al. 2021). This paper aims to further promote methodological pluralism in corpus linguistics by presenting a novel triangulation strategy based on a combination of lexical, move and dialogic analysis. In addition, we showcase a new technique for dialogic analysis which uses the parallel concordance tool to examine conversational patterns. We demonstrate this methodological approach through a case study of a corpus of Twitter interactions involving passengers and airline customer service agents during the first wave of the Covid-19 pandemic. The analysis explores how agents perform *emotional labor* (Hochschild 1983) in their responses to customers complaints. Lexical analysis is used to identify micro-level linguistic devices that are used for expressing and managing emotions, quantitatively assess the prominence of overt emotional displays, and qualitatively investigate the functions emotive linguistic devices serve in the discourse. Move analysis complements lexical analysis by enabling us to account for the totality of pragmatic acts performed and map them onto emotional labor strategies, assess the degree of routinization of emotional labor, and examine the interplay of pragmatic acts via examination of move sequences. Finally, dialogic analysis looks beyond messages produced by a single participant to examine turn exchanges between interactants. This is important because corpus linguistic studies often do not include examination of discourse structure or conversational interaction, focusing on patterns *across* texts rather than patterns *within* texts ('intratextual' analysis, see Caple et al 2020: 27-28). Specifically, in our case study we use dialogic analysis to determine how given emotions expressed by the initiator – the complaining customer – are addressed by the responder – a customer service agent.

References

Baker, P., & Egbert, J. (2016). *Triangulating methodological approaches in corpus linguistic research*. Routledge.

Bednarek, M. (2019). *Creating dialogue for TV: Screenwriters talk television*. Routledge.

Caple, H., Huan, C., & Bednarek, M. (2020). *Multimodal News Analysis across Cultures*. Cambridge University Press.

Fuoli, M., Clarke, I., Wiegand, V., Ziezold, H. & Mahlberg, M. (2021). Responding effectively to customer feedback on Twitter: A mixed methods study of webcare styles. *Applied Linguistics* 42(2), 569–595.

Hochschild, A. (1983). *The Managed Heart: The Commercialization of Human Feeling*. University of California Press, Berkeley, CA.

Establishing a 'new normal': detecting fluctuating trends in word frequency over time

Andrew Kehoe, andrew.kehoe@bcu.ac.uk
Matt Gee, matt.gee@bcu.ac.uk
Antoinette Renouf, antoinette.renouf@bcu.ac.uk
(Birmingham City University)

In this paper we conduct a diachronic study of a corpus covering over 30 years of mainstream UK news text by developing new statistical methods for analysing time series graphs. Such time series analysis has been of growing interest in the study of language change (e.g. Petersen et al. 2012, Grieve et al. 2017). In previous work (Kehoe et al. forthcoming), we presented approaches to finding instances of word frequency change in a data-driven manner. 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. In a follow-up study, we developed a new collocational method for explaining the word frequency changes detected by our data-driven approach. By producing collocational profiles for every word type in the corpus on a month-by-month basis in the form of horizon graphs (Saito et al. 2005), we were able to demonstrate, for example, that the upward trend in gender is related to new collocates such as trans, (non-)binary and fluid.

However, one major issue remained unresolved in our previous work. The trend test as originally formulated was calculated using the entire history of our news corpus which, at that point, was only 10 years. As our corpus has grown to cover 30 years, so too has the possibility of a word exhibiting multiple trends in frequency across the corpus. Reducing the window size for the test provides some improvement, but the test result still lags behind the frequency data, making it difficult to find precise change points. For example, public decreases in frequency from 2011 to 2016, but this is not detected due to the earlier upward trend in public between 2007 and 2010. To be able to detect a downward trend following a pronounced upward trend (or vice versa), we must find an appropriate way of resetting our statistical tests, or establishing a 'new normal' from which subsequent variations can be found.

Our approach is to reframe the issue as one of time-series segmentation, in which our tool attempts to divide the frequency history of a word into time spans exhibiting consistent upward or downward change. We propose a sequence of steps to achieve this: 1) extract the trend from the time series as a moving average using a Kolmogorov–Zurbenko filter (Yang & Zurbenko 2010), 2) segment the time series at peaks and troughs of the trend, based on the gradient changing from positive to negative (or vice versa), 3) test the trend in each resulting segment to filter out minor changes, in this instance we once again apply Cox's sequential test. Regarding public, we find segments with significant trends in the ranges January 2002 – January 2007 (downward), February 2007 – April 2011 (upward) and May 2011 – April 2016 (downward). This process can be further refined by accounting for seasonal variation and identifying abrupt shifts in frequency (as described by Boulton & Lenton 2019).

We conclude by noting further analysis that can build on our approach. The frequency of semantically-related collocates may fluctuate in unison, which opens up the possibility of measuring correlation between time series. For example, during the upward trend in public from February 2007 to April 2011, its collocates finances, spending and cuts show similar change.

References

Boulton, Chris & Timothy Lenton. 2019. A new method for detecting abrupt shifts in time series [version 1; peer review: 2 approved with reservations]. F1000Research, 8:746.

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.

Kehoe, Andrew, Matt Gee & Antoinette Renouf. (forthcoming). A data-driven approach to finding significant changes in language use through time series analysis. In Susanne Flach & Martin Hilpert (eds.) Language in time, time in language, Amsterdam: John Benjamins.

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).

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. IEEE Symposium on Information Visualization INFOVIS 2005, 173-180.

Yang, Wei & Igor Zurbenko. 2010. Kolmogorov–Zurbenko filters. WIREs Computational Statistics, 2(3), 340-351.

Phrasal verbs as multiword units: A comparison of EFL and ESL

Gaëtanelle Gilquin (University of Louvain)
gaetanelle.gilquin@uclouvain.be

Phrasal verbs in English have been described as multiword verbs, because they are made up of two elements, a verb and a particle, but essentially behave like a single verb (see Quirk et al. 1985: 1150). Unlike some other multiword units whose composition is fixed, however, the two elements of the phrasal verb are separated from each other in certain contexts (e.g. 'He filled it up'). This special feature makes the phrasal verb a particularly interesting multiword unit to investigate, especially among non-native speakers of English, since phrasal verbs are said to be "one of the most notoriously challenging aspects of English language instruction" (Gardner & Davies 2007: 339).

This paper centres around phrasal verbs with 'up' in two types of non-native English, namely English as a foreign language (EFL) and English as a second language (ESL), using data from ICLE, LINDSEI and ICE as well as reference corpus data representing English as a native language (ENL). In an earlier study (Gilquin 2015), it was shown that phrasal verbs tend to be underused in both EFL and ESL, but that ESL displays a more native-like stylistic distribution, with phrasal verbs being more frequent in speech than in writing. In this follow-up study, the focus is on the link between the verb (V) and the particle (P), and more precisely on their possible separation by an object (O), resulting in the distinction between VPO and VOP constructions. Two research questions are addressed:

(i) How do EFL and ESL varieties compare with each other with respect to the distribution of phrasal verbs between VPO and VOP constructions?
(ii) Do EFL and ESL users display different preferences in terms of the verbs they combine with 'up' in VPO vs VOP constructions?

It is hypothesized that, because of the higher degree of exposure to English in ESL than in EFL environments (see, e.g., Biewer 2011), phrasal verbs in ESL should be used in a more native-like manner than in EFL.

After extracting all the occurrences of 'up' from the different corpora and manually discarding the instances where 'up' was not part of a phrasal verb, the remaining c. 7,000 phrasal verbs were encoded as to their structure (VPO or VOP), the nature and length of the object (if any), and the verb used. For the second research question, distinctive collexeme analyses were carried out by means of Coll.analysis (Gries 2007).

The results show that the verb and the particle are more often kept together (VPO) in EFL than in ESL, and that the latter is more similar to ENL in that respect. Interestingly, the two non-native varieties appear to use VPO even with very short nominal objects and some pronouns, which is less often the case in ENL. The choice of verbs in both EFL and ESL appears to be more native-like with VPO than with VOP. These results partly confirm the more native-like command of phrasal verbs in ESL than in EFL, and also point to the better entrenchment of VPO than VOP in both non-native varieties. Possible explanations for these findings will be discussed.

References

Biewer, C. 2011. Modal auxiliaries in second language varieties of English: A learner's perspective. In J. Mukherjee & M. Hundt (eds) Exploring second-language varieties

of English and learner Englishes: Bridging a Paradigm Gap (pp. 7-33). Amsterdam: John Benjamins.
Gardner, D. & M. Davies. 2007. Pointing out frequent phrasal verbs: A corpus-based analysis. TESOL Quarterly 41: 339-359.
Gilquin, G. 2015. At the interface of contact linguistics and second language acquisition research: New Englishes and Learner Englishes compared. English World-Wide 36(1): 91-124.
Gries, S. Th. 2007. Coll.analysis 3.2a. A program for R for Windows 2.x.
Quirk, R., S. Greenbaum, G. Leech & J. Svartvik. 1985. A Comprehensive Grammar of the English Language. London: Longman.

Explaining regional patterns in morphosyntactic dialect features: The case of BE *sat/stood* in England and beyond

Jason Grafmiller (University of Birmingham), j.grafmiller@bham.ac.uk
Jack Grieve

In this study we examine a well-known yet little understood dialect feature of English, namely the past participle forms *sat* and *stood* with progressive meaning, as illustrated in (1) and (2).

(1) My work colleague is sat eating honey out of a jar with a spoon.
(2) I was stood chatting with a friend in the street in Darlington when we saw it.

Though frequently mentioned in dialect surveys (e.g. Cheshire, Edwards & Whittle 1989; Kortmann & Lunkenheimer 2014), relatively little is known about this feature's precise geographical distribution or historical origins. Prior studies have relied on relatively small datasets (e.g. Klemola 1999; Stange 2016), which limit our ability to discern reliable synchronic or diachronic patterns. For this study we collected over 150k tokens of the lemma be followed by *sat, sitting, stood*, or *standing* from large-scale corpora, both contemporary (Table 1) and historical (Table 2), to explore the current geographical distribution of BE *sat/stood* and reconsider some competing hypotheses about its origins. Drawing upon this evidence, we argue that BE *sat/stood* is most likely a recent innovation rather than a remnant of an older Germanic BE perfect system, e.g. *I'm not seen it* (cf. Buchstaller & Corrigan 2015).

On Twitter, BE *sat/stood* is widespread throughout most of England and Wales, and is particularly prominent in the North and Southwest of England (Figure 1). Data from the corpus of Global Web-based English (GloWbE) confirms that BE *sat/stood* is quite rare outside the UK (Figure 2; see also Kortmann & Lunkenheimer 2014), and we find no mention of it in the literature of other English varieties, including regions that retain a more productive BE perfect system (Filppula 2004:75; Melchers 2004:39–40; Werner 2016). This is all the more striking in light of well documented founder effects in North America for other (older) morphosyntactic features (e.g. Montgomery 2009; Strelluf 2020; Yerastov 2016). Further, we find only scant traces of BE *sat/stood* in historical corpora prior to the 1700s (cf. Kytö 1997), and no mention of it in pre-20th century dialect surveys or commentaries. Our findings therefore suggest that BE *sat/stood* is not likely an echo of the BE perfect, despite their superficial similarities.

Following Klemola (1999), we propose that BE *sat/stood* is more likely a relatively recent 18th century innovation, and represents a curious case of a change from above resulting in a change from below. It is likely that progressive BE *sat* is a result of a hypercorrection of the older past participle form sitten as it was increasingly replaced by the standard form *sat* in the 17th and 18th centuries. This older variant persisted in the North of England into the 1800s (Klemola 1999), where northern speakers over-extended *sat* to progressive contexts in which sitten was homophonous with the progressive participle variant sittin' [sɪtɪn].

We conclude with a brief look at the linguistic and external factors conditioning variation between *sat* and sitting, and discuss directions for future research.

Table 1: Frequencies of *BE sat/sitting* and *BE stood/standing* in four modern corpora

	sat		*sitting*		*stood*		*standing*	
Twitter	64,785	(59.9%)	43,348	(40.1%)	9,361	(40.2%)	13,931	(59.8%)
GloWbE	2,182	(6.3%)	32,655	(93.7%)	836	(4.1%)	19,498	(95.9%)
Bank of English	468	(3.0%)	15,045	(97.0%)	253	(2.3%)	10,860	(97.7%)
Spoken BNC 2014	273	(33.7%)	537	(66.3%)	66	(27.5%)	174	(72.5%)

Table 2: Frequencies of *BE sat/sitting* and *BE stood/standing* in historical corpora (rounded counts are approximate)

	Period	*sat*	*sitting*	*stood*	*standing*
Early English Books Online (EEBO)	1470-1690	20	1400	50	1000
ARCHER	1600-1999	0	49	0	48
Corpus of Late Modern English Texts (CLMET)	1710-1920	6	950	1	750
Old Bailey Corpus	1720-1913	7	2440	3	4380
Hansard Corpus (British Parliament)	1800-1900	10	2000	0	550
Corpus of Historical American English	1800-pres	13	18000	6	17000

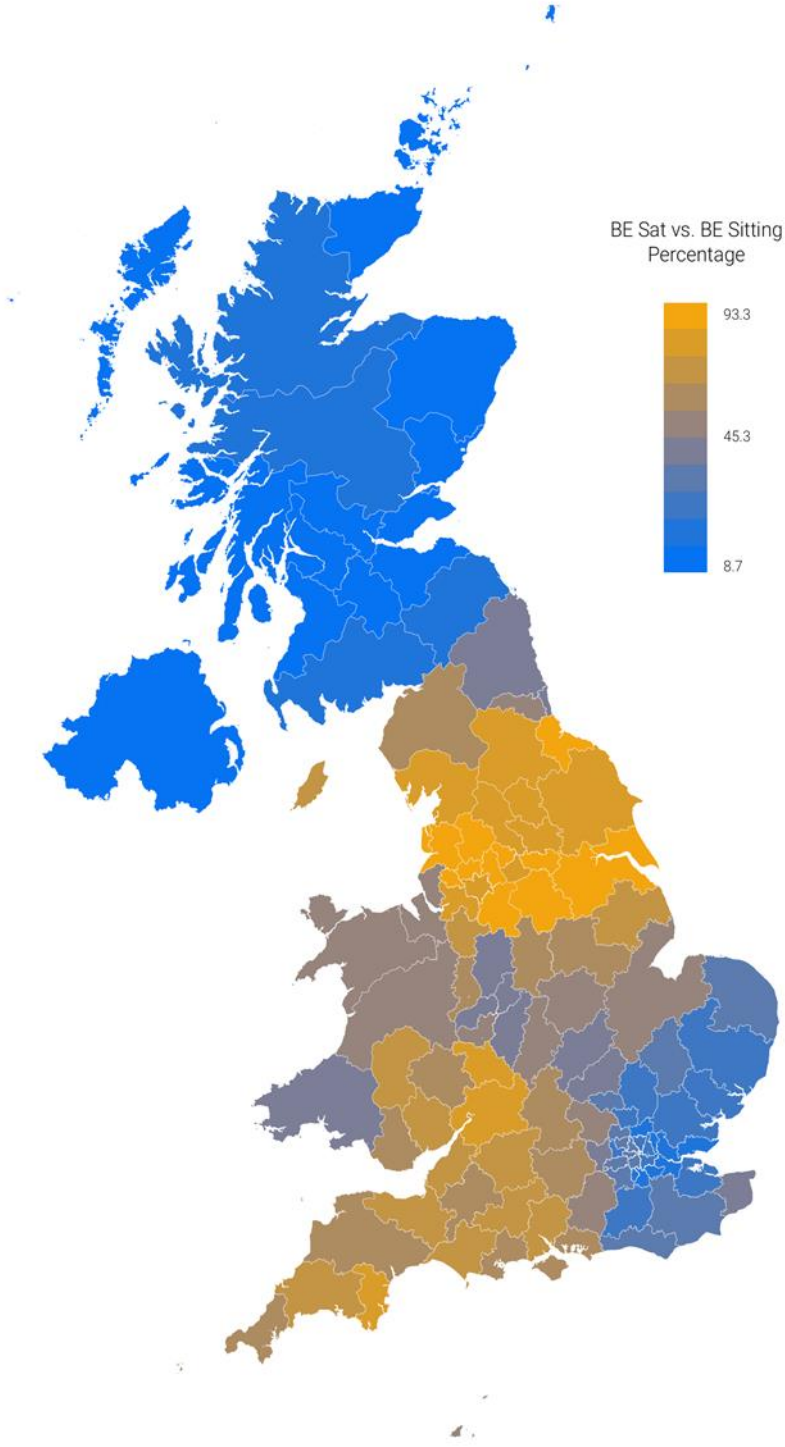

Figure 1: Percentage of BE *sat* vs. *sittng* on UK Twitter 2014 (N*sat* = 64785, N*sitting* = 43348).

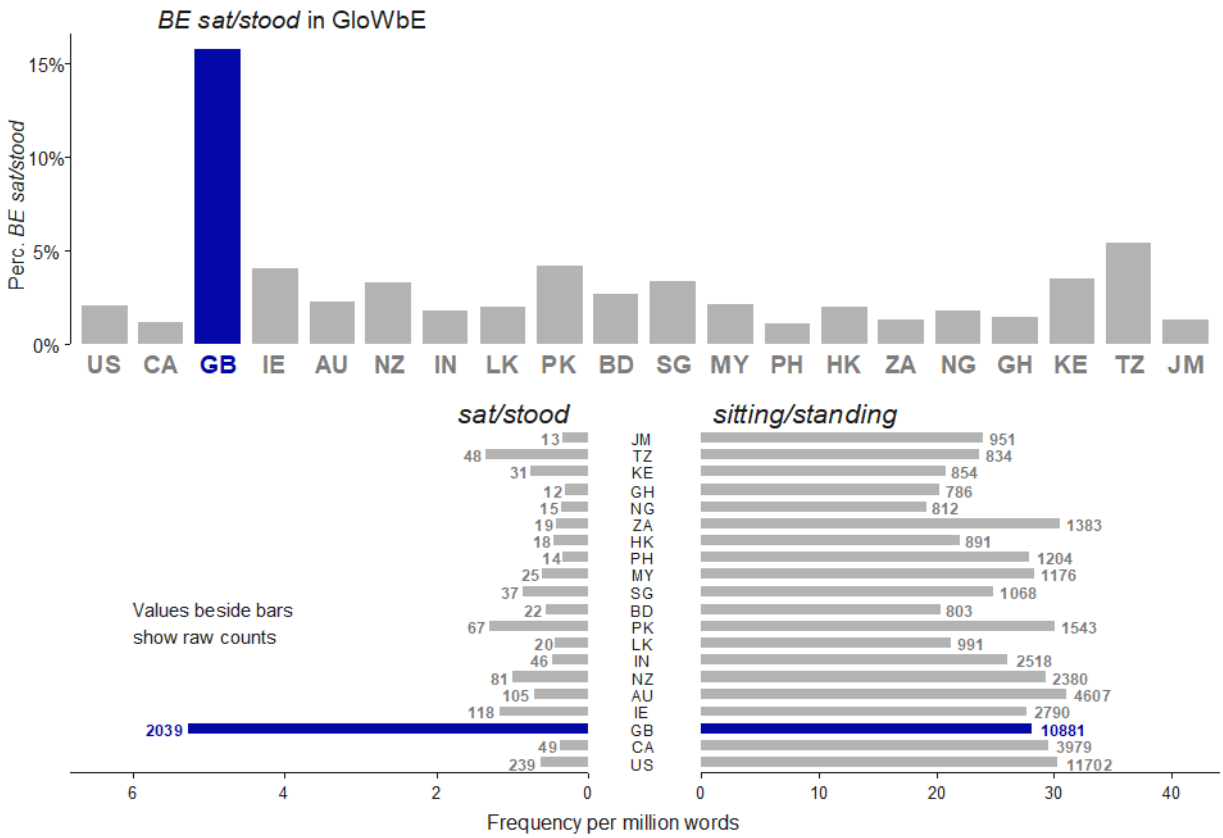

Figure 2: Proportions and requencies of BE sat/stood and BE sitting/standing in GloWbE (N = 55171).

References

Buchstaller, Isabelle & Karen P. Corrigan. 2015. Morphosyntactic features of Northern English. In Raymond Hickey (ed.), Researching Northern English, 71–98. Amsterdam: John Benjamins Publishing Company.

Cheshire, Jenny, Viv K. Edwards & Pamela Whittle. 1989. Urban British Dialect Grammar: The Question of Dialect Levelling. English World-Wide 10(2). John Benjamins. 185–225.

Filppula, Markku. 2004. Irish English: Morphology and syntax. In Bernd Kortmann, Edgar Schneider, Kate Burridge, Raj Mesthrie & Clive Upton (eds.), A Handbook of Varieties of English, vol. 2: Morphology and Syntax, 73–101. Berlin: Mouton de Gruyter.

Klemola, Juhani. 1999. Still sat in your car? Pseudopassives with sat and stood and the history of non-standard varieties of English English. Sociolinguistica 13(1). De Gruyter. 129–140.

Kortmann, Bernd & Kerstin Lunkenheimer. 2014. The electronic world atlas of varieties of English. http://ewave-atlas.org.

Kytö, Merja. 1997. Be/Have + past participle: The choice of the auxiliary with intransitives from Late Middle to Modern English. In Matti Rissanen, Merja Kytö & Kirsi Heikkonen (eds.), English in transition, 17–86. Berlin ; New York: Mouton de Gruyter.

Melchers, Gunnel. 2004. English spoken in Orkney and Shetland: Morphology, syntax, and lexicon. In Bernd Kortmann, Edgar Schneider, Kate Burridge, Raj Mesthrie & Clive

Upton (eds.), A Handbook of Varieties of English, vol. 2, 34–46. Berlin: Mouton de Gruyter.
Montgomery, Michael B. 2009. Historical and comparative perspectives on A-prefixing in the English of Appalachia. American Speech 84(1). 5–26.
Stange, Ulrike. 2016. I was sat there talking all night: A corpus-based study on factors governing intra-dialectal variation in British English 1. English Language & Linguistics 20(3). Cambridge University Press. 511–531.
Strelluf, Christopher. 2020. Needs +PAST PARTICIPLE in regional Englishes on Twitter. World Englishes 39(1). 119–134.
Werner, Valentin. 2016. Rise of the undead? Be-perfects in World Englishes. In Valentin Werner, Elena Seoane & Cristina Suárez Gómez (eds.), Re-assessing the present perfect, 259–294. (Topics in English Linguistics volume 91). Berlin ; Boston: De Gruyter.
Yerastov, Yuri. 2016. Reflexes of the transitive be perfect in Canada and in the US: A comparative corpus study. Dialectologia: revista electrònica. 167–199.

Most dispersion measures do not measure dispersion

Stefan Th. Gries (UCSB & JLU Giessen)
stgries@gmail.com

The two most widely-used corpus statistics by far are probably frequencies (of occurrence and of co-occurrence) and association measures (such as MI and log-likelihood). However, over the last 10 years or so, a variety of publications have also made a case for a more widespread adoption/use of dispersion measures, i.e. measures that quantify the degree to which (typically) words are distributed evenly or 'clumpily' in a corpus (Savický & Hlaváčová 2002; Gries 2008, 2010, 2021; Biber et al. 2016; Burch et al. 2017; Egbert & Biber 2019). While I agree with the notion that dispersion information is important, in this paper, I will do three things.

First, I will argue that nearly all dispersion measures that are currently used do in fact not measure dispersion well – instead, they merely repackage frequency information. To support this seemingly bold/counterintuitive claim, I will discuss results supporting it on the basis of 12 dispersion measures (including Juilland's D, Rosengren's S, KLD, IDF, DP/DPnorm, range) applied to 6 corpora (including the BNC, the BNCspoken, Brown, and the ICE-GB) that show two things:

- most dispersion measures are 0.9 correlated with frequency of occurrence (based on R2s of generalized additive models);
- if frequency and dispersion measures were really measuring different constructs independently of each other, it should be possible to identify words with high and low frequencies with both high and low dispersions, but the way dispersion measures are computed practically rules out findings words that are of low frequency and even dispersion.

Second, I will outline how we can measure dispersion in a way that is truly independent of frequency. I will first use a straightforward example to motivate the proposed way of measuring (two specific words in the Brown corpus), then I will discuss how the measure can be computed and how its computation makes it independent of frequency. Specifically, the new approach involves computing for each frequency of a word type in a corpus the minimally possible dispersion measure (set to 0) and the maximally possible dispersion measure (set to 1) and then determining where on that scale the actually observed dispersion measure for a word falls. I will then apply it to the same 6 corpora on which the traditional measures were tested and show how much less than the traditional measures it is correlated with frequency.

Finally, I will validate the measure on the basis of psycholinguistic data from the Massive Auditory Lexical Decision (MALD) database (Tucker et al. 2019). When the new measure (as applied to the 6 corpora) is compared to existing ones in terms of how well it, together with (logged) frequency as a second dimension, predicts lexical decision times (between 68K and 112K tokens, depending on the corpus used), for 5 out of 6 corpora, it beats all other measures' predictive power.

On the basis of the above results, I will argue that this new measure should be used instead of the traditional ones (at least when 'word commonness' is what is being studied) and I will briefly discuss an additional example of how this measure can also augment collocation/association statistics.

References

Biber, D., Reppen, R., Schnur, E., & Ghanem, R. (2016). On the (non) utility of Juilland's D to measure lexical dispersion in large corpora. International Journal of Corpus Linguistics 21(4): 439–64.

Burch, B., Egbert, J. & Biber, D. (2017). Measuring and interpreting lexical dispersion in corpus linguistics. Journal of Research Design and Statistics in Linguistics and Communication Science 3(2): 189–216.

Egbert, J., & Biber, D. (2019). Incorporating text dispersion into keyword analyses. Corpora 14(1): 77–104.

Gries, St.Th. (2008). Dispersions and adjusted frequencies in corpora. International Journal of Corpus Linguistics 13(4): 403–37.

Gries, St.Th. (2010). Dispersions and adjusted frequencies in corpora: further explorations. In Gries, St.Th., Wulff, St., & Davies, M. (eds.), Corpus linguistic applications: current studies, new directions (pp. 197–212). Amsterdam: Rodopi.

Gries, St.Th. (2021). A new approach to (key) keywords analysis: using frequency, and now also dispersion. Research in Corpus Linguistics 9(2) 1–33.

Savický, P & Hlaváčová, J. (2002). Measures of word commonness. Journal of Quantitative Linguistics 9(3): 215–31.

Tucker, B.V., Brenner, D. Danielson, K. Kelley, M.C., Nenadić, F., & Sims, M. (2019). The Massive Auditory Lexical Decision (MALD) database. Behavior Research Methods 51: 1187–1204.

Many association measures do not measure association (but frequency), and what to do about that

Stefan Th. Gries (UCSB & JLU Giessen), stgries@gmail.com
Magali Paquot

The most widely-used corpus statistics apart from frequencies of occurrence and of co-occurrence are probably association measures (AMs), whose purpose usually is to quantify (i) the degree to which two 'things' like to co-occur, especially collocations for the co-occurrence of words and colligations/collostructions for the co-occurrence of words with syntactic constructions or (ii) the degree to which words are key for a corpus. As is well known, dozens of AMs have been proposed – Pecina (2009) alone reviews more than 80 – but the frequencies of their uses are just as Zipfian-distributed as the words or constructions to which they are applied: a small handful of AMs (arguably, the log-likelihood ratio, $p_{\text{Fisher-Yates exact}}$, *MI* and derivatives such as MI^2 and MI^3, *t*, and Dice) probably covers the vast majority of applications.

In this paper we will demonstrate that most of these association measures actually have a validity problem when one makes the pretty common-sense assumption that a valid measure *M* of *X* is one that measures *X* and not much else (so that *M* can in fact be interpreted in terms of *X* rather than, maybe some quantity of something else). We will demonstrate the AMs do not measure what they are purported to measure (association) – instead, they 'repackage' frequency information and capture very little of association. Thus, in this paper, we will do four things.

First, we will demonstrate on the basis of several hypothetical collostruction results and two keyness results that the maybe most widely used association measure, the log-likelihood ratio, reacts more to frequency than to association; then we will show that the same is true in actual data (Adj-N collocations of *fast* and three other speed adjectives in the British National Corpus). Results from generalized additive models show that the loglikelihood ratio values for the collocations are (i) extremely predictable from the co-occurrence frequency alone (R^2>0.94) and are (ii) hardly correlated at all with the potential gold standard measure (of the log odds ratio) that *only* measures association (R^2<0.06).

Second, for the same data, we will also show that the loglikelihood ratio can be nearly predicted perfectly from an interaction of frequency and association, but with frequency playing the much stronger role: association can only affect the loglikelihood ratio with greater than average frequencies. That high *per definitionem* correlation between frequency and the loglikelihood ratio of course also means that that AM does not permit the user to identify low-frequency-but-high-association collocations …

Third, we will summarily discuss other association measures to show, for instance, that (i) the *t*-score fares even worse in terms of its validity, that (ii) *MI* fares much better, that (iii) MI^2 and MI^3 are very problematic in how they take a good measure (*MI*) and systematically make it worse, and that (iv) Dice and log Dice differ strongly in terms of how much they return assocaition rather than frequency.

Finally, we will outline a three-step procedure of how one can take any AM and decouple it from frequency. We exemplify this approach with the speed adjectives and show that it is indeed uncorrelated with frequency (R2 with co-occurrence frequency < 0.01) but that it is nearly perfectly correlated with what should be the gold standard AM, the log odds ratio. We conclude with some comments on the use of cut-off points for significant/interesting associations and a pointer to how the logic underlying this frequency-less association measure can be used for other corpus statistics.

Are snowclones the new normal? Using corpora to study extravagant formulaic patterns

Stefan Hartmann (University of Düsseldorf) & Tobias Ungerer
hartmast@hhu.de

In line with the conference theme, we investigate the pattern [X BE the new Y], a typical example of 'snowclones', or "schemas that grow from relatively fixed micro-constructions that are usually formulae or clichés" (Traugott & Trousdale 2013: 150). Snowclones have gained increasing interest in recent research on linguistic creativity and in studies on extravagance and expressiveness in language (e.g., Traugott & Trousdale 2013; Bergs 2019; Tizón Couto 2021). So far, however, snowclones have only been coarsely defined, and they have not been studied in a detailed corpus-based fashion.

In our paper, we develop an operational definition of snowclones based on three criteria: (i) the existence of an (alleged) lexically fixed source construction; (ii) partial productivity; (iii) distinctive ('extravagant') formal and/or functional characteristics. We illustrate the three criteria with an in-depth analysis of [X BE the new Y], as in scientists are the new pop stars or data is the new oil.

Drawing on data from COCA (Davies 2008–) and the web corpus ENCOW (Schäfer & Bildhauer 2012), we use collostructional analysis (Stefanowitsch 2013) and distributional semantics (Perek 2016) to explore the typical semantics and the productivity of the two open slots in [X BE the new Y]. We show that the snowclone is generally productive, but that its slot fillers are still centered around specific semantic domains, such as colors (pink is the new black), media-related concepts (blogs are the new resume) and group membership terms (Democrats are the new Conservatives). Moreover, our analysis sheds light on the relationship between the X and the Y slot in the construction. We find that [X BE the new Y] is often used to explain abstract X concepts in terms of more concrete Y concepts (Twitter is the new cigarette); that some of the most typical X-Y combinations consist of (near-)antonyms (small is the new big); and that the snowclone frequently encodes innovative, non-trivial comparisons (truth is the new hate speech). Together, the partial productivity of [X BE the new Y] and the fact that it tends to express unusual and creative comparisons indicate that the snowclone is an 'extravagant' pattern in Haspelmath's (1999) sense. Finally, we discuss the relationship between snowclones and linguistic creativity, focusing on the question of what social, cultural and interpersonal factors influence speakers' choice of salient linguistic constructions.

In sum, we argue that the concept of snowclones, if properly defined and analyzed with state-of-the-art corpus tools, can contribute substantially to our understanding of creative language use. In particular, such analyses can shed light on the question of how social, cultural, and interpersonal factors combine to influence the choice of more or less salient linguistic constructions.

References

Bergs, Alexander. 2019. What, if anything, is linguistic creativity? Gestalt Theory 41(2). 173–183.

Davies, Mark. 2008. The Corpus of Contemporary American English (COCA). https://www.english-corpora.org/coca/.

Haspelmath, Martin. 1999. Why is grammaticalization irreversible? Linguistics, 37(6), 1043–1068.

Perek, Florent. 2016. Using distributional semantics to study syntactic productivity in

diachrony. A case study. Linguistics 54(1). 149–188.
Schäfer, Roland & Felix Bildhauer. 2012. Building large corpora from the web using a new efficient tool chain. In Nicoletta Calzolari, Khalid Choukri, Terry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), Proceedings of LREC 2012, 486–493.
Stefanowitsch, Anatol. 2013. Collostructional analysis. In Thomas Hoffmann & Graeme Trousdale (eds.), The Oxford Handbook of Construction Grammar, 290–306. Oxford: Oxford University Press.
Tizón Couto, David. 2021. The rise of COVID 'snowclones' – the mother of all linguistic phrases. https://theconversation.com/the-rise-of-covid-snowclones-the-mother-of-all-linguistic-phrases-167580.
Traugott, Elizabeth Closs & Graeme Trousdale. 2013. Constructionalization and constructional changes. (Oxford Studies in Diachronic and Historical Linguistics 6). Oxford: Oxford University Press.

Exploring emerging patterns of self-identification in the *LGBTQ+ Reddit Corpus*

Turo Hiltunen, turo.hiltunen@helsinki.fi
Laura Hekanaho
Minna Palander-Collin
Helmiina Hotti

In recent years, our understanding of both gender and sexuality has broadened considerably, allowing room for various identities to emerge even in mainstream discourse. Yet the starting place of many recent changes has often been the gender and sexuality minorities themselves. These social developments have also led to new ways of talking about identities, particularly in terms of self-identification practices (e.g. Diamond et al. 2011, Galupo et al. 2015). For example, the acknowledgment of transgender and nonbinary individuals has reshaped the public discussion on gender, bringing questions about a person's right to self-identify to the center instead of biological or external criteria. The concurrent emergence of online registers as productive sites for identity-related discussions has established social media sites used by the LGBTQ+ community as a fruitful data source for exploring new practices of sexual and gender identity formation, and, more broadly, the link between social developments and language practices (cf. Baker 2014, Loureiro-Porto & Hiltunen 2020).

While some attention has been given to identity labels (e.g. White et al. 2018, Zimman 2017), our work-in-progress report focuses on constructions used for self-identification on the popular discussion forum *Reddit*, in particular *identify as* N, *as a* N, and *be* N. Through analysis of lexicogrammatical patterns we specifically investigate in what ways and to what extent such constructions are employed to construct minority gender and sexual identities, and how salient a feature self-identification is in these fora. Along with identification, these constructions are also employed to position oneself in discourse, and as such, they reveal broader trends in discursive identity practices.

In order to explore the emergence of self-identification patterns, we have collected *The Reddit LGBTQ+ Corpus* (c. 44 million words), including discussions about minority gender and sexualities from a plethora of LGBTQ+ related subforums on *Reddit* (www.reddit.com; e.g., r/lgbt, r/nonbinary, r/bisexual). The corpus covers the time period from January 2010 to November 2021. The corpus contains approximately 600 submissions per month and their subsequent comments extracted from the *Pushshift* repository (e.g. Baumgartner et al. 2020).

Our preliminary findings suggest that the three constructions in focus are productive in the corpus as rhetorical means for claiming a specific identity (e.g. *I identify as non-binary most days*). At the same time, these constructions are often used for labelling others, together with meta-discussion on the appropriate demarcation of these categories. In the paper, we illustrate the usage of these constructions and contextualise them with respect to the ongoing discourse and positioning. We also reflect on the potential of this new dataset to generate more nuanced questions for further study.

References

Baker, Paul. 2014. *Using Corpora to Analyse Gender*. London: Bloomsbury.
Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M. & Blackburn, J. 2020. *The Pushshift Reddit Dataset*. Available from: https://arxiv.org/abs/2001.08435

Diamond, L., Pardo, S. & Butterworth, M. 2011. Transgender experience and identity. In Schwartz S., Luyckx, K. & Vignoles, V. (Eds) *Handbook of identity theory and research*, 629–647. New York: Springer.

Galupo, M., Mitchell, R. & Davis, K. 2015. Sexual minority self-identification: Multiple identities and complexity. *Psychology of Sexual Orientation and Gender Diversity*, 2(4), 355–364.

Loureiro-Porto, L. & Hiltunen T. 2020. "Democratization and Gender-neutrality in English(es)". *Journal of English Linguistics* 48(3): 215-232. (Special issue: *Democratization and Gender-neutrality in English(es)* ed. by Loureiro-Porto L & Hiltunen T.) doi:10.1177/0075424220935967

White, A., Moeller, J., Ivcevic, Z., & Brackett, M. 2018. Gender identity and sexual identity labels used by U.S. high school students: A co-occurrence network analysis. *Psychology of Sexual Orientation and Gender Diversity, 5*(2), 243–252.

Zimman, L. 2017. Transgender language reform: some challenges and strategies for promoting trans-affirming, gender-inclusive language. *Journal of Language and Discrimination*, 1(1).

Rhythm in World Englishes – Evidence from a Quantitative Analysis of Co-occurrence Patterns in Corpora of L1 and L2 Varieties of English

Sebastian Hoffmann (Trier University), hoffmann@uni-trier.de
Sabine Arndt-Lappe (Trier University), arndtlappe@uni-trier.de
Peter Uhrig (FAU Erlangen-Nürnberg), peter.uhrig@fau.de

This paper investigates the connection between stress and rhythm in World English. More specifically, it attempts to test the hypothesis that the impact of rhythmically different L1-contexts can be measured in the (written) output of institutionalised second-language varieties of English.

It is a well-established fact that languages have rhythmic properties. Following Pike (1945) and Abercrombie (1965, 1967), languages have traditionally been categorised as stress-timed (e.g. English) or syllable-timed (e.g. Spanish); in addition, a number of languages have been classified as being mora-timed (e.g. Japanese, see e.g. Han 1962). More recent experimental research (e.g. Dauer 1983, 1987) has shown that these rhythmic classes are not clearly defined and that we are instead dealing with a continuum of rhythmic variation.

For English, there is a considerable body of research on what has been termed the Principle of Rhythmic Alternation ('PRA', Sweet 1876) – i.e. the general tendency to maintain an alternation of stressed and unstressed syllables. The bulk of this research is on written data (or on orthographically transcribed speech) and focuses on preferences in lexical or grammatical choice (e.g. *drúnken sáilor* instead of *drúnk sáilor*) or word ordering preferences (e.g. *compléte and únabridged* instead of *únabridged and compléte*) that are interpreted as resulting from stress-clash – or stress-lapse – avoidance strategies (see. e.g. Schlüter 2005; Shih 2017). Complementing this work, there is a growing body of corpus-based research in phonology assessing the status of metrical constraints on a more global scale. Based on simple bigram probabilities in a large variety of corpora comprising more that 10 million words, Breiss & Hayes (2020) show that metrically critical bigrams – i.e. phonetic contexts deemed less preferable by the PRA – are underrepresented in their data.

Post-colonial varieties of English are typically claimed to exhibit clear tendencies towards syllable timing (see e.g. the list in Mesthrie & Bhatt 2008: 129). For Singapore English, for example, this classification is supported in studies by Low and colleagues (e.g. Low 1998, Low & Grabe 1995, Low et al. 2000); for a book-length study of speech rhythm in acrolectal Indian English, see Fuchs (2016). All inner-circle varieties of English (cf. Kachru 1985), however, are said to be stress-timed. Given the difference between inner and outer circle varieties of English, the PRA should therefore apply to different degrees, since rhythmic well-formedness is less likely to play a prominent role in most – if not all – L2-varieties. As we will demonstrate, this hypothesis can indeed be confirmed by applying the method used by Breiss & Hayes to data from the 20 components of GloWbE (Davies & Fuchs 2015).

In a second step, we will then explore the potential of the method for highlighting individual processes of structural nativization in World Englishes. In particular, we will be asking the question whether a focus on rhythmic well-formedness – or, in fact, the lack thereof – can offer relevant pointers for the detection of structures that would otherwise remain under the radar of researchers.

References

Abercrombie, D. (1965) *Studies in Phonetics and Linguistics.* London: Oxford University Press.

Abercrombie, D. (1967) *Elements of General Phonetics.* Edinburgh: Edinburgh University Press.

Breiss, C. & B. Hayes (2020) 'Phonological markedness effects in sentence formation', *Language* 96(2). 338–370.

Dauer, R.M. (1983) 'Stress-timing and syllable-timing re-analysed', *Journal of Phonetics*, 11: 51–62.

Dauer, R.M. (1987) 'Phonetic and phonological components of language rhythm', *Proceedings of the 11th International Congress of Phonetic Sciences*, 447–50.

Davies, M. & Fuchs, R. (2015) 'Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English Corpus (GloWbE)', *English World-Wide* 36(1): 1–28.

Fuchs, R. (2016) *Speech Rhythm in Varieties of English: Evidence from Educated Indian English and British English.* Singapore: Springer.

Han, M.S. (1962) 'The feature of duration in Japanese', *Onsei no kenkyuu*, 10: 65–80.

Kachru, B.B. (1985) 'Standards, codification, and sociolinguistic realism: The English language in the outer circle.' In: Quirk, R & Widdowson, H.G. (eds.), *English in the World: Teaching and Learning the language and the literature*. Cambridge: Cambridge University Press. 11–30.

Low, E.L. (1998) *Prosodic prominence in Singapore English*, PhD thesis (University of Cambridge).

Low, E.L. & Grabe, E. (1995) 'Prosodic patterns in Singapore English', *Proceedings of the 13th International Congress for Phonetic Sciences*, 636–9.

Low, E.L., Grabe, E. & Nolan, F. (2000) 'Quantitative characterizations of speech rhythm: syllable-timing in Singapore English', *Language and Speech*, 43, 4: 377–401.

Mesthrie, R. & Bhatt, R.M. (2008) *World Englishes*. Cambridge: Cambridge University Press.

Pike, K.L. (1945) *The Intonation of American English* (Ann Arbor: University of Michigan Press).

Schlüter, J. (2005) *Rhythmic Grammar. The Influence of Rhythm on Grammatical Variation and Change in English.* Berlin & New York: Mouton de Gruyter.

Shih, St. S. (2017) 'Phonological Influences in Syntactic Alternations.' In: Gribanova, V. & St.S. Shih (eds.), *The Morphosyntax-Phonology Connection*, Oxford: Oxford University Press. 223–252.

Sweet, H. (1876) 'Words, Logic, and Grammar', *Transactions of the Philological Society, 1875-1876*. 470–503.

Promotion and Preservation of Public Health: Trends in Health Science and Communication in the Royal Society Corpus

Katherine Ireland (University of Georgia)
katherine.ireland@uga.edu

This proposed presentation investigates ongoing changes in linguistic patterns and discourses in medical and scientific communication in the first English scientific periodical, the Philosophical Transactions of the Royal Society of London (Phil Trans), utilizing the Royal Society Corpus (Fischer et al. 2020). The Royal Society Corpus (RSC) is composed of the Phil Trans from its beginning in 1665 to 1920 (Fischer et al. 2020); it is approximately 78 million words and has been encoded for text types, year of publication, and tokenized and linguistically annotated for lemma and part of speech using the TreeTagger (Schmid 1995).

Efforts in public health have been traced back to scientific advances from the beginning of the Royal Society (RS), and many distinguished members of the RS made significant contributions to these efforts (Wootton 2015; Berridge 2016). From its beginnings, the Phil Trans provided a new means for "disseminating the scientific information that provided momentum to the scientific movement that still continues today," with emergent and distinctive genre characteristics (Kronick 1990; Atkinson 1999: xxii; Gross et al. 2002: viii; Kermes et al. 2016; Biber & Conrad 2019: 222; Biber and Conrad 2019: 236-7).

In the 1650s, authors of the Phil Trans describe fever as "Nature's Engine". This simple phrase demonstrates key departures from early natural philosophy to a greater understanding of processes in living organisms. It also represents a crossover between physical and life sciences and is a notable example of the interaction between cultural norms, scientific developments, and medical practice. As shown in previous studies (Ireland Kuiper forthcoming; Atkinson 1999; Biber and Conrad 2019; Gotti 2011; Monaco 2016; Tang and Rundblad 2017; Biber and Conrad 2019), fundamental changes in science align with linguistic changes in communication. This proposed presentation expands on previous research on the Phil Trans by focusing on the consideration of specific tokens and keywords, including health, disease(s), and inoculate|inoculation using collocational and concordance analysis to understand important discourse trends. R packages polmineR (Blätte & Leonhardt 2019) and ggplot (Wickham, Navarro, and Pederson 2021) are implemented for analysis and visualizations of the data. Key findings include departures from Early Modern perspectives of health and medicine and the impact of cultural and scientific developments on understandings of health and disease (Wootton 2015: 21-22). A distinctive focus on medical practice and public health also increases over time in collocations and keywords surrounding the tokens of interest.

References

Atkinson, Dwight. 1999. Scientific Discourse in Sociohistorical Context: The Philosophical Transactions of the Royal Society of London 1675-1975. Lawrence Erlbaum Associates.

Baker, Paul. 2006. Using Corpora in Discourse Analysis. London: Continuum.

Berridge, Virginia. 2016. Public Health: A Very Short Introduction. Oxford University Press.

Biber, Douglas, and Susan Conrad. 2019. Register, Genre, and Style: 2nd Edition. Cambridge, UK: Cambridge University Press.

Blätte, A. and C. Leonhardt. 2019. PolmineR() package, v 0.8.0.
Brezina, V. 2018. Statistics in Corpus Linguistics: A Practical Guide. Cambridge University Press.
Collins, Luke and Brigitte Nerlich. 2016. Uncertainty discourses in the context of climate change: A corpus-assisted analysis of UK national newspaper articles. Communications. De Gruyter Mouton. 291-313.
Evert, S. and A. Hardie. 2011. 'Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium', in Proceedings of the Corpus Linguistics 2011 conference, University of Birmingham, UK.
Evert, Stefan, and the CWB Development Team. 2021. The IMS Open Corpus Workbench (CWB) Corpus Encoding and Management Manual. http://cwb.sourceforge.net/
Gläser, Rosemarie. 1995. Linguistic Features and Genre Profiles of Scientific English. Berlin: Peter Lang.
Gotti, Maurizio. 2011. The development of specialized discourse in the Philosophical Transactions. In Medical Writing in Early Modern English. (Eds. Irma Taavitsainen and Päivi Pahta). Cambridge University Press.
Gross, Alan, Joseph Harmon, and Michael Reidy. 2002. Communicating Science: The Scientific Article from the 17th Century to the Present. Oxford, UK: Oxford University Press.
Hardie, A. 2012. 'CQP Web: combining power, flexibility, and usability in a corpus analysis tool', International Journal of Corpus Linguistics, pp. 380-409. John Benjamins Publishing.
Kermes, Hannah, Stefania Degaetano-Ortleib, Ashraf Khamis, Jörg Knappen, and Elke Teich. 2015. The Royal Society Corpus: From Uncharted Data to Corpus.
Kretzschmar, W., C. Darwin, C. Brown, D. Rubin, D. Biber. 2004. Looking for the Smoking Gun: Principled Sampling in Creating the Tobacco Industry Documents Corpus. Journal of English Linguistics. 32:1.
Monaco, Leida. 2016. Was late Modern English scientific writing impersonal? Comparing Philosophy and Life Sciences texts from the Coruña Corpus. International Journal of Corpus Linguistics. John Benjamins Publishing. 499-526.
R Core Team 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria.
Tang, Chris, and Rundblad, Gabriella. 2017. When Safe Means 'Dangerous': A Corpus Investigation of
Risk Communication in the Media. Applied Linguistics. Oxford University Press.Wickham, Hadley, Danielle Navarro, and Thomas Lin Pederson. 2021. ggplot2: Elegant Graphics for Data Analysis. https://ggplot2-book.org/index.html
Wootton, David. 2015. The Invention of Science: A New History of the Scientific Revolution. Harper Publishing.

Grammatical Nativization in Spoken South Asian Englishes:
The Case of the Existential-There Construction

Kathrin Kircili (University of Marburg), kathrin.kircili@uni-marburg.de
Julia Degenhardt, Julia.Degenhardt@admin.uni-giessen.de
Tobias Bernaisch
Sandra Goetz

In Schneider's Dynamic Model of Postcolonial Englishes, the stage of nativization encompasses a "restructuring of the English language" (Schneider 2007: 44) and thus the manifestation of linguistic innovations on all linguistic levels. These linguistic developments represent the structural independence of a given variety from its historical input variety – by default British English (BrE) – accompanied by the gradual establishment of a pan-ethnic identity within the speech community concerned.

Research into variety-specific structures of South Asian Englishes has predominantly focused on Indian English (IE) as the largest and sociolinguistically most evolved South Asian variety (cf. e.g. Mukherjee 2007; Sedlatschek 2009; Bernaisch 2015) while the remaining varieties have not received comparable scholarly attention. A case in point is Sri Lankan English (SLE), where, so far, mainly lexical (cf. e.g. Meyler 2007; Bernaisch 2015), lexico-grammatical (cf. e.g. Bernaisch et al. 2014; Gries & Bernaisch 2016) and pragmatic (cf. e.g. Kraaz & Bernaisch 2020; Bernaisch 2022) nativization processes have been studied. Notably, syntactic studies on SLE and South Asian Englishes more generally – particularly those that adopt statistically multifactorial approaches – are rare and have – to the best of our knowledge – exclusively studied written texts (cf. e.g. Bernaisch 2015; Götz 2017).

Against this backdrop, this paper investigates the existential-*there* construction in spoken South Asian Englishes (viz. Indian and Sri Lankan Englishes) in comparison to BrE. Previous research points to the fact that the construction, in which the expletive *there* occupies the position of the grammatical subject with the notional subject being moved to post-verbal (i.e. post-BE) position (*There*[Sg] *is a book.*[Sn]), has already undergone a nativization process in IE. The result is the *Indian* or *non-initial* existential, in which the expletive and the notional subject are reversed (*A book is there.*) (cf. Balasubramanian 2009; Lange 2012; Winkle 2015).

Hence, the present study aims at answering the following research questions:

1) Can quantitative and qualitative differences between the three varieties be attested for existential *there*?

2) Which predictors guide structural choices of existentials in both South Asian Englishes as well as BrE and to what extent are the effects of said predictors cross-varietally distinct?

3) Is the structural evidence of existential *there* compatible with the more advanced evolutionary status of IE compared to SLE?

To answer these research questions, 4,500 existentials from the spoken parts of the British, Indian and Sri Lankan components of the *International Corpus of English* (ICE) were analyzed manually and annotated for more than 20 predictors including, for example, sociolinguistic (e.g. GENDER of the speaker) and structural (e.g. LENGTH of the notional subject) ones. Tree-based models such as a generalized linear mixed-model tree

(cf. Fokkema et al. 2020) are expected to depict a complex interplay of these predictors, accounting for the more frequent use of the non-initial structure in IE (9.98%) compared to SLE (1.02%) and BrE (0.22%) (χ^2=374.26, df=2, *p*-value<0.0001, Cramer's *V*=0.232), which appears compatible with the difference in evolutionary progress between IE and SLE.

References

Balasubramanian, C. (2009): Register Variation in Indian English. Amsterdam: John Benjamins.

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

Bernaisch, T. (2022): "Features of Sri Lankan English", English in East and South Asia: Policy, Features and Language in Use, ed. E.L. Low & A. Pakir. London: Routledge. 168–182.

Bernaisch, T., S.Th. Gries & J. Mukherjee (2014): "The dative alternation in South Asian Englishes: Modelling predictors and predicting prototypes", English World-Wide 35(1), 7–31.

Fokkema, M., J. Edbrooke-Childs & M. Wolpert (2020): "Generalized linear mixed-model (GLMM) trees: A flexible decision-tree method for multilevel and longitudinal data", Psychotherapy Research 31(3), 329–341.

Götz, S. (2017): "Non-canonical syntax in south-Asian varieties of English: A corpus-based pilot study on fronting", ZAA - Zeitschrift für Anglistik und Amerikanistik 65(3), 265–281.

Gries, S.Th. & T. Bernaisch (2016): "Exploring epicentres empirically: Focus on South Asian Englishes", English World-Wide 37(1), 1–25.

Kraaz, M. & T. Bernaisch (2020): "Backchannels and the pragmatics of South Asian Englishes", World Englishes [Early view].

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

Meyler, M. (2007): A Dictionary of Sri Lankan English. Colombo: Mirisgala.

Mukherjee. J. (2007): "Steady states in the evolution of New Englishes: Present-day Indian English as an equilibrium", Journal of English Linguistics 35(2), 157–187.

Schneider, E.W. (2007): Postcolonial English: Varieties around the World. Cambridge: Cambridge University Press.

Sedlatschek, A. (2009): Contemporary Indian English: Variation and Change. Amsterdam: John Benjamins.

Winkle, C. (2015): Non-Canonical Structures, They Use Them Differently: Information Packaging in Spoken Varieties of English. Dissertation, Albert Ludwigs University Freiburg.

A corpus-based acoustic analysis of vowel production by L1-Japanese learners and native speakers of English

Yuki Komiya (The University of Queensland), y.komiya@uqconnect.edu.au
Martin Schweinberger

This study combines acoustic phonetics, (applied) corpus linguistics, machine learning, and speech recognition to analyse the production of the monophthongal vowels /ɪ iː e ʌ æ ɑ ʊ uː/ in the speech of L1-Japanese learners and L1-speakers of English using transcripts and audio data from the Japanese spoken monologue section of The International Corpus Network of Asian Learners of English (ICNALE). The ICNALE is a multi-modal international learner corpus representing more than 10,000 topic-controlled speeches and essays produced by college students from China, Hong Kong, Indonesia, Japan, Korea, Pakistan, the Philippines, Singapore/Malaysia, Taiwan, and Thailand as well as English native speakers and their socio-demographic and language proficiency information.

In contrast to English, where vowel pairs differ in both duration and vowel space position, e.g., /ɪ iː/ and /ʌ ɑ/ and /ʊ uː/, it has been claimed that Japanese vowels differ only in duration but not in vowel space position (see Kubozono 2015). The aim of this acoustic analysis is to evaluate what vowels L1-Japanese learners struggle with in terms of target-like vowel production and to provide insights into the determining factors causing divergencies from L1-English produced vowels. Target-like production, or target-proximity, is operationalized in the form of Pillai overlap scores as well as Euclidean distance. We hypothesize that L1-Japanese learners will produce target-proximate vowels when producing vowels that are either very similar or very distinct (or non-existing) in Japanese and English. In contrast, Japanese learners are expected to deviate most strongly from target-like vowel production when the English vowel has a similar, yet different equivalent in Japanese. In that case, we expect to observe (near) mergers of /ɪ iː/ and /ʌ ɑ/ and /ʊ uː/ (see Flege et al. 2003). In addition, we hypothesize that learners will exaggerate the duration of long vowels of spectrally similar vowels (/ɪ iː/ or /ʌ ɑ/ or /ʊ uː/) to compensate for the lack of differentiation among (near)-mergers.

The transcripts and audio-files of 300 L1 Japanese learners and 300 L1 English speakers were aligned and combined into Praat TextGrids using WebMAUS Basic (Kisler et al. 2017, Schiel 1999). F1, F2, and F3 formants and durations of vowels were extracted from the TextGrids in R (R Core Team). In the analysis, American English served as target variety leading us to exclude all non-North American L1 speakers of English. As the quality of the recordings led to substantive noise interference, outliers were removed based on kernel density estimation with only tokens with density values in the upper .8 percentile being retained. To statistically test our hypotheses, we used Bhattacharyya distance and Pillai scores to determine if vowels were merged. In addition, we used a MuPDARF approach (see Gries & Deshors 2014) to determine how and where L1 and L2 speakers differed in their vowel production.

The results of the analysis confirm that L1-Japanese learners of English exhibit higher degrees of overlap, i.e., indicators of merging for /ɪ iː/ and /ʌ ɑ/ and /ʊ uː/. The distances between these (near-)mergers are significantly shorter compared to their L1-English peers. With respect to duration, the analysis shows that L1-Japanese learners do indeed exaggerate the duration of long vowels to compensate for the lack of qualitative differences between short and long vowel pairs.

This study innovative in that in that it is the first corpus-based study which analyses acoustic traits of learner-produced vowels by applying a machine-learning approach to a large collection of learners' spontaneous speech. The results can be used to

raise awareness of L1-specific difficulties among this learner cohort due to their L1-background. As such, this study showcases how the application of speech recognition and machine-learning as well as the extension of corpus linguistics to acoustic phonetics and applied linguistics could potentially inform the development of targeted teaching materials for learners with specific L1-backgrounds.

References

Flege, J. E., Schirru, C., & MacKay, I. R. A. (2003). Interaction between the native and second language phonetic subsystem. Speech Communication 40(4), 467-491.

Gries, S. T. & Deshors, S. C. (2014). Using regressions to explore deviations between corpus data and a standard/target: two suggestions. Corpora 9(1), 109-136.

Ishikawa, S. (2014). Design of the ICNALE Spoken: A new database for multi-modal contrastive interlanguage analysis. Learner Corpus Studies in Asia and the World 2: 63-76.

Kisler, Th., U. D. Reichel, & F. Schiel (2017). Multilingual processing of speech via web services. Computer Speech & Language 45: 326–347.

Kubozono, H. (2015). Introduction to Japanese phonetics and phonology. In H. Kubozono (Ed.), Handbook of Japanese phonetics and phonology, 1-40. De Gruyter.

Schiel, F. (1999). Automatic Phonetic Transcription of Non-Prompted Speech. In Proceedings of the International Congress of Phonetic Sciences: 607-610.

A complex puzzle: Comparing theory-based models of grammatical complexity in spoken versus written registers

Tove Larsson (NAU - Northern Arizona University),tove.larsson@nau.edu
Douglas Biber
Gregory R. Hancock

Over the last several decades, numerous studies have compared the differing grammatical complexities of spoken and written registers. These studies have repeatedly found that spoken registers rely on clausal complexity features, whereas written registers rely on phrasal complexity features (see, e.g., Biber et al., 2022). One less noticed finding from previous research, however, is that the spoken and written modes differ fundamentally in the extent to which the use of complexity features can be manipulated, in that the spoken registers "are produced and comprehended in real-time, setting a cognitive ceiling for the syntactic and lexical complexity typically found in these [registers]" (Biber, 1988:163). This fundamental difference is described in even greater detail in Biber (1992):

> [W]ritten registers differ widely among themselves in both the extent and kinds of discourse complexity, while spoken registers follow a single pattern with respect to their kinds of complexity, differing only with respect to extent (p. 159).

In the present paper, we return to these claimed fundamental differences between the discourse complexities of speech and writing. In particular, we test the two major claims made in Biber (1992):

1. For any given complexity feature/parameter, is there more variability across registers in the written mode than in the spoken mode?
2. Do all complexity features/parameters follow a single 'pattern' (i.e., covarying in the same ways) across spoken registers, versus multiple patterns of variation across written registers?

Our examination is based on analysis of multiple complexity features in a corpus of spoken and written registers that are matched pairwise across the modes in terms of level of interactivity, level of expertise of the audience, and communicative purpose (e.g., conversational opinions vs. opinion blogs; classroom teaching vs. textbooks).

We applied structural equation modeling (SEM) techniques to assess the adequacy of these two claims; specifically, we used mean structure models and confirmatory factor models. In the traditional, non-SEM application of inferential statistics, the typical hypothesis is simply the claim that an association is not due to random chance – a very weak claim. In contrast, in SEM techniques, the hypotheses are specific models proposed on the basis of previous research and theory, and the statistical analysis allows comparisons across multiple models to determine which best accounts for the data. Larsson et al. (2021) called for greater use of SEM techniques in corpus-based research, because they require a much deeper engagement with the claimed generalizable findings from previous research.

Early results show that a model with equal variation across registers in the two modes has worse fit than a model that permits greater variation across registers in the written mode, supporting the claim that there is more variability in the use of complexity features across registers in the written mode. Confirmatory factor techniques are used to compare the adequacy of models specified such that all complexity features function as

part of a single underlying mechanism in the spoken versus written modes, supporting the claim that all complexity feature variation follows a single pattern in speech but not in writing.

References

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press.

Biber, D. (1992). On the complexity of discourse complexity: A multidimensional analysis. *Discourse Processes*, 15, 133–163.

Biber, D., Gray, B., Staples, S., & Egbert, J. (2022). *The register-functional approach to grammatical complexity: Theoretical foundation, descriptive research findings, application*. Routledge.

Larsson, T., Plonsky, L., & Hancock, G. R. (2021). On the benefits of structural equation modeling for corpus linguists. *Corpus Linguistics and Linguistic Theory*, 17(3), 683–714.

Adverb placement in L2 spoken production: The effect of linguistic and extralinguistic factors

Tove Larsson (NAU- Northern Arizona University),tove.larsson@nau.edu
Marcus Callies, Tülay Dixon
Hilde Hasselgård (University of Oslo), hilde.hasselgard@ilos.uio.no
Nicole Hober, Natalia Judith Laso, Isabel Verdaguer, Sanne van Vuuren, Magali Paquot

Despite being a rich source of information about potential syntactic first-language (L1) transfer, studies of adverb placement in second-language (L2) users' production are few and far between (though notable exceptions include Rankin, 2010, Hasselgård, 2015, and Larsson et al., 2020). What is more, existing research has tended to focus on misplaced adverbs in written data and rarely considered factors such as mode, adverb type, and the linguistic context surrounding the adverb of interest. The present study starts out where previous studies left off and looks at adverb placement in intermediate to advanced learners' spoken language production, as outlined below.

In Larsson et al.'s (2020) study on adverb placement in academic writing, only minor differences were noticed across L1 and L2 production, which led the authors to conclude that while certain traces of L1 transfer persisted, L2 writers seem to have largely mastered adverb placement in English. However, the production circumstances of formal writing are such that there is ample opportunity for pre-planning and post-editing of the text, while online spoken language production offers no or very limited opportunities of this sort. It is thus reasonable to expect that effects of L1 transfer might still be discernable in spoken data. Furthermore, Larsson et al. (2020) found linguistic features (e.g., subject type, presence/absence of auxiliary) to be more important predictors of adverb placement in writing than extralinguistic factors (e.g., L1 background). It is unclear whether we can expect these patterns to hold true also in speech. Against this background, the present study sets out to explore two research questions:

- What extralinguistic (e.g., L1 background) and linguistic factors (e.g., verb type, presence/absence of auxiliaries) help predict the positional distribution of the adverbs investigated in the spoken production of L1 and L2 students?
- What evidence (if any) of L1 transfer is found in spoken language production of advanced learners of English?

Specifically, we studied the positional distribution of 15 epistemic adverbs (e.g., *maybe*, *apparently*, *actually*; Granath, 2002) in spoken L1 English data from LOCNEC and in L2 data from six L1 backgrounds (French, Norwegian, Swedish, Dutch, German, and Spanish) from LINDSEI (Gilquin, De Cock, & Granger, 2010). The study used Hasselgård's (2010) syntactic classification of adverbs, the analytical framework developed in Larsson et al. (2020; available at www.iris-database.org), and Jarvis's (2000) framework for transfer identification. Our 4,002 tokens were coded manually by two trained raters, and we used Random Forests to examine the data. The results show that similarly to Larsson et al.'s findings for the written data, the linguistic variables proved to be the best predictors, with adverb type (e.g., *clearly*, *importantly*) and presence/absence of an auxiliary verb being particularly important. Only very limited traces of L1 transfer were found (e.g., instances of SVAO word order for speakers of languages that allow verb raising, i.e., French and Spanish). The present study helps broaden our understanding of the interplay between linguistic and extralinguistic factors and the production constraints of spoken production, thus providing one more piece of the puzzle that makes up adverb placement.

References

Gilquin, G., De Cock, S., & Granger, S. (Eds.). (2010). LINDSEI: Louvain International Database of Spoken English Interlanguage. UCL Presses.

Granath, S. (2002). The position of the adverb certainly will make a difference. *English Today*, 18(1), 25–30.

Hasselgård, H. (2015). Lexicogrammatical features of adverbs in advanced learner English. *International Journal of Applied Linguistics*, 166(1), 163–189.

Hasselgård, H. (2010). *Adjunct Adverbials in English*. Cambridge University Press.

Jarvis S. (2000). Methodological rigor in the study of transfer: identifying L1 influence in the interlanguage lexicon. *Language Learning*, 50(2), 245–309.

Larsson, T., Callies, M., Hasselgård, H., Laso, N. J., Van Vuuren, S., Verdaguer, I., & Paquot, M. (2020). Adverb placement in EFL academic writing: Going beyond syntactic transfer. *International Journal of Corpus Linguistics*, 25(2), 155–184.

Rankin, T. (2010). Advanced learner corpora data and grammar teaching: Adverb placement. In M.C. Campoy, B. Belles-Fortuno, & M.L. GeaValor (Eds.), *Corpus-based Approaches to English Language Teaching* (pp. 205–215). Continuum.

'Guv, I'm a copper, not a social worker!': Using corpus-assisted discourse studies to analyse how caring professionals are portrayed on anglophone TV

Maria Leedham (The Open University)
Maria.leedham@open.ac.uk

Many professionals in broadly-categorised 'caring' domains feel they are poorly-represented in television dramas (e.g. Chatterjee, 2020; Weaver et al., 2013). In the case of social workers, previous research has indicated that both press and entertainment media consistently portray the profession negatively, particularly in child protection cases (e.g. Reid & Misener, 2001; Zugazaga et al 2006). This study builds on previous work on the portrayal of social workers in UK press articles (Leedham, 2021), extending this scope to consider how jobs broadly classified as 'caring' are portrayed in English-medium TV programmes first shown in the period 2010-2017, and seeks to answer the following research questions:

1) What are the proportions of positive, negative and neutral mentions of *social worker(s)* on TV?
2) How does this portrayal differ from that of other caring professionals?
3) What is the nature of the negativity around *social worker*?

The research takes a social constructionist approach to knowledge creation; methodologically, the study combines corpus linguistics with discourse analysis in exploring how the language surrounding mentions of professionals constructs, reinforces and extends the wider societal view of the profession (cf. studies in Taylor and Marchi, 2018). The dataset employed is the 325 million word *TV Corpus* (Davies, 2019) comprising transcripts from a broad array of TV dramas from anglophone contexts; this is explored through collocate lists and concordance lines (n=1600) from eight subcorpora featuring the professions of *social worker, nanny, teacher, doctor, cop, therapist, priest* and *nurse*. Two hundred concordance lines per profession were categorised using five levels from highly positive to highly negative by two independent raters, and negative categories were then further coded. Additionally, IMDB* programme plot synopses featuring *social worker* were explored to investigate the positioning of this professional group within programmes. Findings indicate a highly negative portrayal of *social workers* as either judgmental bureaucrats or uncaring childsnatchers, and also suggests that social worker characters on TV – in common with those from other female-dominated professions such as *nanny* and *nurse* – are frequently-portrayed as having inappropriate sexual relationships with clients. Insights into how different professionals are portrayed were also revealed through the search term *n*t a social worker* - as in the title quotation - wherein characters seek to distance themselves from particular professions perhaps viewed as lower status.

The study contrasts with previous research on how professions are portrayed on TV as the focus here is on the *language* surrounding mentions of the professionals rather than their visual depictions or characterisation through actions. As such, it exemplifies the use of corpus linguistic procedures alongside more qualitative methods and illustrates the widespread use of corpora across research areas as part of the 'new normal' of corpus research. The study furthers understanding of the ways in which social workers and other professionals are portrayed in television dramas through the dialogue of characters, illustrating the role of entertainment media in shaping widely-held views on different professionals.

References

* Internet Movie DataBase https://www.imdb.com/

Chatterjee, Deepshikha, and Ann Marie Ryan. 2020. 'Is policing becoming a tainted profession? Media, public perceptions, and implications', Journal of Organizational Behavior, 41: 606-21.

Davies, Mark. (2019) *The TV Corpus*. Available online at https://www.english-corpora.org/tv/.

Leedham, M. (2021). ''Social Workers Failed to Heed Warnings': A Text-Based Study of How a Profession is Portrayed in UK Newspapers', The British Journal of Social Work. https://doi.org/10.1093/bjsw/bcab096

Reid, W. J., & Misener, E. (2001). Social work in the press: a cross-national study. International Journal of Social Welfare, 10(3), 194-201.

Taylor, C., & Marchi, A. (2018). *Corpus approaches to discourse: A critical review*. London and New York: Routledge.

Weaver, Roslyn, Yenna Salamonson, Jane Koch, and Debra Jackson. 2013. 'Nursing on television: student perceptions of television's role in public image, recruitment and education', Journal of Advanced Nursing, 69: 2635-43.

Zugazaga, C.B., R.B. Surette, M. Mendez, and C.W. Otto. 2006. 'Social worker perceptions of the portrayal of the profession in the news and entertainment media: an exploratory study', Journal of Social Work Education, 42: 621.

"I shall be glad if you will note…" – Studying early 20th century business correspondence from Hong Kong to assess variety-specific genre developments

Lisa Lehnen (University of Würzburg), lisa.lehnen@uni-wuerzburg.de
Ninja Schulz (University of Würzburg), ninja.schulz@uni-wuerzburg.de
Carolin Biewer

In world Englishes research, it is necessary to clearly delineate genres and trace their evolution in the specific context because genre might be a stronger predictor of variation than variety (Noël & van der Auwera, 2015). Ideally, diachronic corpora of postcolonial varieties should be representative of different genres at different periods. Unfortunately, the text types preserved are not necessarily the most fitting because they are very formal, dominated by native British speakers, and thus less likely to show structural nativisation. Still, the language may reflect local usage that left an imprint on the genre or the variety as a whole. In the mid-19th century, banks and companies, such as the Hong Kong Shanghai Banking Corporation (HSBC) and Jardine Matheson & Co., established their head offices in Hong Kong, so that business correspondence has been in place from early onwards. Although it was initially dominated by writers whose native language was British English of the late 19th century, we argue that this is the input which should serve as a starting point to analyse variety-specific developments, since business was one of the domains in which British and local staff interacted (Hao, 1982, pp. 86–87; HSBC, n.d.; Life at HSCB, 2016) and British speakers in Hong Kong can be considered the "expert speakers" of the time (Kirkpatrick, 2007, as cited in Groves, 2012).

In this study, we explore the language of early business correspondence in Hong Kong by looking at small-sized corpus (ca. 20,000 words) containing 125 letters written in the 1930s. To illustrate instances of genre- and potentially variety-specific language, we focus on (i) lexical borrowing, (ii) archaisms, and (iii) modal verbs. According to Schneider's Dynamic Model (2007, 39, 55), lexical borrowing for cultural terms, customs and objects starts during the second phase. Examples from our corpus, e.g., *cumshaw*, *tiffin*, *gampei*, reflect instances of borrowing from the local language into English. Vice versa, archaisms, such as *mephitic* or *thereto*, and formal language used in the letters may have been acquired by locals and integrated into their usage even beyond the domain of business. Modal verbs are similarly sensitive to cultural impacts (Biewer et al., 2020; Leech, 2013) and of paramount importance in business correspondence as stance and politeness markers (Del Lungo Camiciotti, 2006a, 2006b; Dossena, 2006a, 2006b) and part of legal language (Dossena, 2010). To identify their functions in business correspondence from Hong Kong, we provide an analysis of their frequency and usage patterns in our corpus and compare these with the distribution of modal verbs in British English (Leech, 2013; Leech & Smith, 2009) and newspaper writing in Hong Kong (Biewer et al., 2020) of the same time. While this study illustrates single points of interest in a small, specialised corpus, the aim of the larger project is to advance the description of postcolonial varieties by creating a more solid basis for describing variety-specific genre developments even in highly conventionalised and formal genres.

References

Biewer, C., Lehnen, L., & Schulz, N. (2020). "The future elected government should fully represent the interests of Hongkong people" – Diachronic change in the use of modalising expressions in Hong Kong English between 1928 and 2018. In P. Hohaus

& R. Schulze (Eds.), *Re-assessing modal expressions: Categories, co-text, and context* (pp. 311–341). John Benjamins.

Del Lungo Camiciotti, G. (2006a). "Conduct yourself towards all persons on every occasion with civility and in a wise and prudent manner; this will render you esteemed": Stance features in nineteenth-century business letters. In M. Dossena & S. M. Fitzmaurice (Eds.), *Business and official correspondence: Historical investigations* (pp. 153–174). Lang.

Del Lungo Camiciotti, G. (2006b). From Your obedient humble servants to Yours faithfully: The negotiation of professional roles in the commercial correspondence of the second half of the nineteenth century. In M. Dossena & I. Taavitsainen (Eds.), *Diachronic perspectives on domain-specific English* (pp. 153–172). Lang.

Dossena, M. (2006a). Forms of self-representation in nineteenth-century business letters. In M. Dossena & I. Taavitsainen (Eds.), *Diachronic perspectives on domain-specific English* (pp. 173–190). Lang.

Dossena, M. (2006b). Stance and authority in nineteenth-century bank correspondence - a case study. In M. Dossena & S. M. Fitzmaurice (Eds.), *Business and official correspondence: Historical investigations* (pp. 175–192). Lang.

Dossena, M. (2010). "We beg to suggest": Features of legal English in Late Modern business letters. In N. Brownlees, G. Del Lungo Camiciotti, & J. Denton (Eds.), *The Language of Public and Private Communication in a Historical Perspective* (pp. 46–64). Cambridge Scholars Publishing.

Groves, J. M. (2012). The issue of representativeness in Hong Kong English. *Asian Englishes*, *15*(1), 28–45. https://doi.org/10.1080/13488678.2012.10801318

Hao, Y. (1982). The Compradors. In M. Keswick (Ed.), *The thistle and the jade: A celebration of 150 years of Jardine, Matheson & Co* (pp. 85–101). Octopus Books.

HSBC. (n.d.). *Our history: Local staff, local knowledge*. Retrieved May 18, 2020, from https://www.hsbc.com/who-we-are/our-history

Leech, G. (2013). Where have all the modals gone? An essay on the declining frequency of core modal auxiliaries in recent standard English. In J. I. Marín Arrese (Ed.), *English modality: Core, periphery and evidentiality* (pp. 95–115). De Gruyter Mouton.

Leech, G., & Smith, N. (2009). Change and constancy in linguistic change: How grammatical usage in written English evolved in the period 1931-1991. In A. Renouf & A. Kehoe (Eds.), *Corpus linguistics: Refinements and reassessments* (pp. 173–200). Rodopi.

Life at HSCB. (2016). *150 Years - Chapter Three: Local Staff, Local Knowledge* [Video]. YouTube. https://www.youtube.com/watch?v=hOAj_NdpGdo

Noël, D., & van der Auwera, J. (2015). Recent quantitative changes in the use of modals and quasi-modals in the Hong Kong, British and American printed press. In P. Collins (Ed.), *Grammatical Change in English World-Wide* (pp. 437–464). John Benjamins.

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

Semantic reanalysis and idiomatization:
multi-word verbs in the Late Modern English period

Ljubica Leone (Lancaster University)
l.leone1@lancaster.ac.uk

The present study aims to examine the semantic changes affecting multi-word verbs (hereafter MWVs), namely phrasal verbs, prepositional verbs, and phrasal-prepositional verbs, during the Late Modern English (LModE) period. Specifically, the objective is to describe the role performed by semantic reanalysis and idiomatization in the linguistic renewal of MWVs during the years 1750-1850.

Many studies have highlighted that over time interacting factors created the conditions for the grammaticalization of independent units and their lexicalization as verbs characterized by a complex internal constituency (Denison 1981; Claridge 2000). The already formed MWVs often underwent semantic reanalysis leading to the renewal of extant meanings and idiomatization which, since early periods, favored increasing semantic opacity (Denison 1981; Claridge 2000; Elenbaas 2007). However, despite extensive knowledge about the development of MWVs, some areas require further investigation.

Existing works, indeed, suffer from two major limitations: (i) they mostly examined earlier periods in the history of English including the Old English (OE), Middle English (ME), and Early Modern English (EModE) periods (Denison 1981; Hiltunen 1983; Claridge 2000; Elenbaas 2007); (ii) when the focus has been on the more recent LModE period, they have especially investigated phrasal verbs (Brinton 1988; Thim 2012; Leone 2016, 2019; Rodríguez-Puente 2019). This means that the description of the semantic renewal of MWVs during the LModE time remains to date unexplored.

The present study aims to fill this gap and to contribute to existing knowledge about the diachronic development of MWVs. Specifically, the aims are: (1) to describe the semantic changes affecting MWVs during the years 1750-1850; (2) to study the role performed by semantic reanalysis intended as the process leading to semantic extension and/or pragmatic specialization (Eckardt 2006); (3) to examine processes of idiomatization favoring internal demotivation.

The present research is a corpus-based investigation undertaken on the Late Modern English-Old Bailey Corpus (LModE-OBC), a corpus that has been compiled by selecting texts from the Proceedings of the Old Bailey (https://www.oldbaileyonline.org/), London's Central Criminal court. The corpus includes trials and witness depositions dating back to the years 1750-1850 and overall amounts to 1,008,234 words. MWVs have been examined with the software WordSmith Tools 6.0 (https://www.lexically.net/wordsmith/). MWVs were retrieved using the tool 'Concord' that allows concordance-based analysis of selected instances and the visualization of their immediate context.

The analysis reveals that during the years 1750-1850 MWVs underwent processes of semantic renewal which resulted in the creation of new meanings or in the reinterpretation of extant nuances. Specifically, the most important aspects are: (1) phrasal verbs were affected by semantic reanalysis, which favored semantic extension and pragmatic specialization, and by idiomatization; (2) both prepositional verbs and phrasal-prepositional verbs were involved in processes of idiomatization resulting in increasing internal demotivation.

References

Brinton, Laurel J. 1988. The development of English aspectual systems. Aspectualizers and post-verbal particles. Cambridge: Cambridge University Press.

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. Accessed at http://www.escholar.manchester.ac.uk/uk-ac-man-scw:74782

Eckardt, Regine. 2006. Meaning change in grammaticalization. An enquiry into semantic analysis. Oxford & New York: Oxford University Press.

Elenbaas, Marion. 2007. The synchronic and diachronic syntax of the English verb-particle combination. Utrecht: LOT.

Hiltunen, Risto. 1983. 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.

Leone, Ljubica. 2016. Phrasal verbs and analogical generalization in Late Modern Spoken English. ICAME Journal. 40(1): 39-62.

Leone, Ljubica. 2019. Context-Induced reinterpretation of phraseological verbs. Phrasal verbs in Late Modern English. In G.C. Pastor and R. Mitkov (eds.), EUROPHRAS 2019, LNAI 1175, 253-267. Cham: Springer Nature.

Rodríguez-Puente, Paula. 2019. The English phrasal verb, 1650-present. History, stylistic drifts, and lexicalization. Cambridge: Cambridge University Press.

The Proceedings of the Old Bailey. n.d. Accessed at https://www.oldbaileyonline.org

Thim, Stefan. 2012. Phrasal verbs. The English verb-particle construction and its history. Berlin & Boston: Walter de Gruyter Mouton.

WordSmith Tools. Version 6.0. n.d. Accessed at https://www.lexically.net/wordsmith/

Grammaticalization of Aspect in German and its diachronic parallels in English

Zlata Liwschin (Leibniz University of Hannover)
zlata.liwschin@germanistik.uni-hannover.de

German is commonly not viewed as a typical aspect language, for in German the perfective and imperfective aspectual distinctions are not marked morphologically on the verb, as it is common in the Slavic languages, particularly in Russian (Comrie, 1976; Forsyth, 1970; Leiss, 1992). However, aspect is a grammatical category that is currently being discussed as grammaticalizing in German (Gárgyán, 2014; Krause, 2002). While in Old English and Old High German participial constructions with beon/wesan and sin/wesan respectively exhibited the primary function of creating internal temporal constituency (Reimann, 1997), upholding the aspectual function and gradually extending their combinability with certain verb classes, in their subsequent stages the initial similarities between the two languages developed apart: In Early Modern English the progressive form gradually became an obligatory member in the English verbal paradigm, whereas in Early New High German the durative function of the construction eventually ceased to exist, this development leading to the disappearance of the form in the German language after the 15th century (Reichmann & Wegera, 1993).

In view of the historical developments, the present work endeavours to define the similarities between the two languages German and English in the way they grammaticalize the category of aspect by acquiring progressive aspect forms. While the English progressive is fully grammaticalized, German progressive constructions are lagging behind, but – as stated by Reimann (1997) (and many others) - German is on its way towards developing obligatory, i.e. fully grammaticalized progressive aspect marking. The study strives to uncover the similarities and differences between the progressives in the two languages in the way they emerged. It will be investigated whether the grammaticalization process in the English language resembles the supposedly presently emerging, presumably similar process in Modern Standard German. For instance, a striking parallel exists between the German am- and beim-progressives and the Early Modern English locative constructions of the type 'be in hunting', built also with the prepositions 'on', 'at', or 'upon' (Núñez-Pertejo, 2004), showing a close formal parallel to the Modern German prepositional progressive forms.

To this end, a comparative corpus study of the progressive in Early/Late Modern English as well as of the progressive forms in Present-Day German is conducted that draws on grammaticalization theory (Lehmann, 2015; Diewald & Smirnova, 2012) as well as on aspectual theory (Comrie, 1976; Leiss, 1992; Bache, 1985). For the English part, the current version of the ARCHER corpus is used, and for the investigation of Present-Day German, DWDS corpus data are analyzed, with a focus on conceptually near-spoken register.

The analysis of corpus data indicates that the internal temporal constituency of a situation is increasingly expressed obligatorily by the am-Progressive of the type "Gitarrenmusik ist am aussterben". Furthermore, my research shows that this obligatory expression can particularly be associated with certain lexical aspect classes. In my ongoing research, I investigate the extent to which the am-Progressive behaves syntactically as well as semantically similar to the English Progressive before its complete grammaticalization in the Late Modern English period, such that it may be concluded that both Germanic languages within their relevant stages of diachronic development undergo or underwent a very similar process of grammaticalization of progressive markers, yet at different times in their individual histories.

References

Bache, C. (1985). Verbal Aspect. A General Theory and its Application to Present-Day English. Odense: Odense University Press.

Comrie, B. (1976). Aspect: An Introduction to the Study of Verbal Aspect and Related Problems. Cambridge: Cambridge University Press.

Diewald, G., & Smirnova, E. (2012). Paradigmatic integration: the fourth stage in an expanded grammaticalization scenario. In: K. Davidse, T. Breban, L. Brems & T. Mortelmans (Eds.), Grammaticalization and Language Change. New reflections. Amsterdam: Benjamins, 111-133.

Forsyth, J. (1970). A Grammar of Aspect: Usage and Meaning in the Russian Verb. Cambridge: Cambridge University Press.

Gárgyán, G. (2014). Der am-Progressiv im heutigen Deutsch: Neue Erkenntnisse mit besonderer Hinsicht auf die Sprachgeschichte, die Aspektualität und den kontrastiven Vergleich mit dem Ungarischen. Frankfurt am Main: Peter Lang Edition.

Krause, O. (2002). Progressiv im Deutschen: Eine empirische Untersuchung im Kontrast mit Niederländisch und Englisch. Tübingen: Niemeyer.

Lehmann, C. (2015). Thoughts on grammaticalization. Berlin: Language Science Press.

Leiss, E. (1992). Die Verbalkategorien des Deutschen: Ein Beitrag zur Theorie der sprachlichen Kategorisierung. Berlin: Walter de Gruyter.

Núñez-Pertejo, P. (2004). The progressive in the history of English with special reference to the Early Modern English period: A corpus-based study. München: LINCOM Europa.

Reichmann, O., & Wegera, K. (1993). Frühneuhochdeutsche Grammatik (=Sammlung kurzer Grammatiken germanischer Dialekte. A. Hauptreihe Nr. 12). Tübingen: Niemeyer.

Reimann, A. (1997). Die Verlaufsform im Deutschen und Englischen. Entwickelt das Deutsche eine „progressive form“? Bamberg: Dissertation.

"Like English is used everywhere" – The functions and use of discourse marker like in UAE English

Eliane Lorenz (Justus Liebig University Giessen, Norwegian University of Science and Technology)
eliane.lorenz@anglistik.uni-giessen.de

The current study investigates the use of English as a Lingua Franca (ELF) in the United Arab Emirates (UAE), focusing on the functions and use of the discourse marker like among university students. It is set in Sharjah, one of UAE's seven sovereign emirates, a metropolitan area characterized by intense language contact due to recent, large-scale immigration (Parra-Guinaldo & Lanteigne 2021; Pacione 2005; Siemund et al. 2021). To date, there is a lack of research investigating the use of ELF in the UAE and its status as a new English variety (Siemund et al. 2021). The discourse marker like has received much scholarly attention (e.g., D'Arcy 2017; Diskin 2017; Fuller 2003; Schweinberger 2014). However, it has mainly been studied in native Englishes, and considerably less research focuses on non-native speakers of English (Diskin 2017) or ELF varieties.

The current study addresses these research gaps by employing a small-size spoken corpus consisting of semi-structured interviews, approximately 30 minutes each, conducted with 58 university students in the UAE (word tokens: 139,630). The participants come from a variety of linguistic backgrounds and include both Emirati as well as non-Emirati population. The interviews were conducted as part of a larger project on Language Attitudes and Repertoires in the Emirates (LARES 2019–2021). The spoken data are complemented by a comprehensive online questionnaire. With this unique data source, it is possible to investigate the use of like and to correlate it with different social (non-linguistic, attitudinal) variables.

The study sets out to answer three research questions:

- Do the UAE students show high individual variation as has been argued to be a characteristic of ELF users (e.g., Mauranen 2017)?
- Does this study find support for an assumed accelerated language change in ELF contexts (e.g., Laitinen 2020)?
- Are there functional differences in discourse maker like uses and if yes, can these be explained with the social background of the students?

First results show that like is the third most frequently used word in the interviewees' utterances (n=3,937), with 2,951 (75%) uses as a discourse marker and 986 (25%) other uses. The mean frequency per 1,000 words (ptw) across the entire corpus is 19.5 (median: 16.0), lending support to an accelerated language change in this ELF setting. Yet, the relatively high standard deviation of 14.75 shows that the individual variation among the speakers is comparably large. The lowest frequency is 0.51 ptw and the highest is 55.14 ptw. This is in line with Mauranen (2017) who argued for variability in ELF encounters. A regression analysis shows that the social background of the speakers (gender, citizenship, L1, year of birth, number of languages, college, self-assessed proficiency in English, and the English usage score) cannot explain the variability identified in the use of the discourse maker like. An additional (functional) coding (i.e., co-occurrence with hesitation; sentence position, i.e., clause-initial, medial, final, and non-clausal (see Schweinberger 2014)), will further assess the use of like among the UAE students.

References

D'Arcy, A. (2017). Discourse-pragmatic variation in context. Eight hundred years of LIKE. Amsterdam: Benjamins.

Diskin, C. (2017). The use of the discourse-pragmatic marker 'like' by native and non-native speakers of English in Ireland. Journal of Pragmatics 120, 144–157.

Fuller, J. M. (2003). Use of the discourse marker like in interviews. Journal of Sociolinguistics 7(3), 365–377.

Laitinen, M. (2020). Empirical perspectives on English as a Lingua Franca (ELF) grammar. World Englishes 39(3), 427–442.

Mauranen, A. (2017). A glimpse of ELF. In M. Filppula, J. Klemola, A. Mauranen & S. Vetchinnikova (eds.), Changing English. Global and local perspectives, 223–253. Berlin: De Gruyter Mouton.

Pacione, M. (2005). Dubai. Cities 22(3), 255–265.

Parra-Guinaldo, V., & Lanteigne, B. (2021). Morpho-syntactic features of transactional ELF in Du-bai/Sharjah. In P. Siemund & J. R. E. Leimgruber (eds.), Multilingual global cities: Singapore, Hong Kong, Dubai, 303–320. Singapore: Routledge.

Schweinberger, M. (2014). The discourse marker LIKE: A corpus-based analysis of selected varieties of English. Doctoral dissertation. University of Hamburg.

Siemund, P. Al-Issa, A., & Leimgruber, F. (2021). Multilingualism and the role of English in the United Arab Emirates. World Englishes 40(2), 191–204.

Challenges in deriving a new COLT from the Spoken BNC2014: the case of teenage swearing

Robbie Love (Aston University) & Anna-Brita Stenström
r.love@aston.ac.uk

The Bergen Corpus of London Teenage Language (COLT) (Stenström & Breivik, 1993) is a half-million-word spoken corpus derived from recordings of teenage speakers gathered in the early 1990s, and it is part of the Spoken BNC1994 (BNC Consortium, 2007). This paper describes recent efforts to derive a new sub-corpus of south-east England teenage language from the Spoken BNC2014 (Love et al., 2017) on a post hoc basis, comparable to the original COLT. We then evaluate the utility of COLT 2 by presenting a case study into teenage swearing, building upon other corpus-based studies of swearing in English (McEnery, 2005; Stenström, 2006; Love, 2021) and teenage language more broadly (Stenström et al., 2002). Swearing is an appropriate basis for comparison between the corpora, as it is a common and important part of human communication which is known to be especially frequent in the speech of adolescents (and young adults).

We pose the following research questions:

- What are the design criteria of the original COLT that would need to be present in a new COLT in order to achieve comparability?
- Are there enough texts in the Spoken BNC2014 that meet these criteria so as to build a comparable sub-corpus?
- What are the differences between the corpora in the use of swearing among teenagers?

The design criteria identified in the original COLT are: teenage speakers, from London, holding conversations predominantly with other teenagers (with minimal or no input from speakers of other ages). In attempting to replicate these criteria using texts from the Spoken BNC2014, we initially identified a total of 54 teenage speakers in the corpus. However, according to the metadata, only seven of these speakers were born in London. In order to increase the size of the sub-corpus, we decided to broaden the inclusion criteria to allow speakers from a larger area of south-east England. In total, we identified 15 teenage speakers from the south-east of England who participated in a total of 35 teenager-only conversations across 25 hours of recordings. These texts were isolated to form the new COLT 2 sub-corpus of c. 300,000 tokens, slightly smaller than the original COLT.

The creation of a new COLT sub-corpus presents several challenges, which raise interesting questions about representativeness and transcription practices. We discuss these issues in light of a case study into swearing among teenagers, the findings of which show similar but interestingly not identical trends to those identified in research into swearing among all speakers in the Spoken British National Corpora. For example, while the relative frequency of *fuck* (the most common swear word) is stable across the full spoken corpora, it is significantly lower in COLT 2 when compared to the original COLT.

References

BNC Consortium. (2007). *The British National Corpus, XML Edition.* Oxford Text Archive. http://hdl.handle.net/20.500.12024/2554

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics, 22*(3), *Special Issue: 'Compiling and analysing the Spoken British National Corpus 2014',* 319-344. DOI: 10.1075/ijcl.22.3.02lov

Love, R. (2021). Swearing in informal spoken English: 1990s – 2010s. *Text and Talk, 41, Special Issue: 'Corpus Linguistics across the Generations: In Memory of Geoffrey Leech'*. DOI: 10.1515/text-2020-0051

Stenström, Anna-Brita. 2006. "Taboo words in teenage talk". In Language Variation and Change: Historical and contemporary perspectives. Special issue of Spanish in Context 3:1 (2006), Mar-Molinero, Clare and Miranda Stewart (eds.), 115–138.

Stenström, A-B & Breivik, L.E. 1993. The Bergen Corpus of Teenage Talk. ICAME Journal 17, 128.

Stenström, A-B, Andersen, G., & Hasund, I. K. (2002). *Trends in teenage talk: corpus compilation, analysis and findings*. John Benjamins.

This essay argues that: Connecting metadiscursive nouns and rhetorical moves in research article abstracts

Liu Luda (Jilin University)

In order to gain further insight into the writing of research article abstracts and, in particular, academic interactions in this promotion-oriented genre, this paper explores metadiscursive nouns in the rhetorical moves of research article abstracts in two closely-related but rather distinct disciplines: literary studies and applied linguistics. Using an adopted version of Hyland's (2000) model of the rhetorical structure of abstracts, all the 291 abstracts from each discipline were annotated for rhetorical moves. Following this, the pattern "determiner + noun", the most frequent structure containing metadiscursive nouns (Jiang & Hyland, 2017), was examined in each of the move sub-corpora. Results revealed significant variation in how the literary scholars and the linguists organize their abstracts, specifically in 1) the employment of disciplinary metadiscursive nouns, 2) how arguments are packaged and presented to achieve cohesion in different rhetorical moves, and 3) how the discipline-specific discursive conventions are reflected in both group of scholars. This study offers empirical support for Hyland's (2002) claim in favor of specificity in EAP/ESP instruction and therefore recommends that the English department should address the disciplinary needs of these two neighboring disciplines.

The avoidance of repetition in translation: A multifactorial study of repeated reporting verbs in the Italian translation of the Harry Potter series

Lorenzo Mastropierro

Repetition is a ubiquitous linguistic feature that enacts a variety of functions in a wide range of registers and discourses. In the discipline of stylistics, repetition is seen as playing a significant role in the creation of foregrounding and stylistic effects, especially in literary language (Wales 2011: 366). However, despite its functional and stylistic importance, repetition is systematically avoided in translation, in favour of lexical variety. Ben-Ari (1998: 3) has famously equated this tendency to a universal norm in translation, and a number of corpus-based studies have provided evidence of this trend (e.g. Čermáková & Fárová 2010, Čermáková 2015, 2018, Čermáková & Mahlberg 2018, Mastropierro 2020). The existing literature on the topic though focuses on the investigation of the stylistic effects that repetition avoidance can have in translation, or on the strategies used by translators to avoid repetition. These studies are extremely important in highlighting the manipulative potentials of translators' choices and their impact on the reception of a text, but shed little light on the nature of repetition avoidance in itself. Aiming to redress this gap, this paper focuses on the description of the avoidance of repetition in translation, rather than on its effects, providing a picture of the linguistic context in which such a phenomenon can occur. Through an exploration of the translation of repeated reporting verbs in the *Harry Potter* series in Italian, this paper applies a multifactorial approach to investigate whether and to what extent four factors, representing linguistic features of the source text items, have an effect on the reproduction of repetition in translation or its avoidance. The factors are (i) the frequency, (ii) the number of possible translation equivalents, (iii) the number of different meanings, and (iv) the semantic category of the source text verbs. Results show that the frequency and the semantic type of reporting verb have a significant effect on the likelihood of seeing the repetition of that verb avoided or maintained in translation, while the number of different meanings and translation equivalents do not have a significant effect. By providing a data-based and multidimensional description of repetition in translation in the context of reporting verbs, this paper furthers our understanding of the phenomenon, with potential implications for translation training and professional practice.

References

Ben-Ari, N. (1998). The ambivalent case of repetitions in literary translation. Avoiding repetitions: A 'universal' of translation. *Meta*, 43(1): 68-78.

Čermáková, A. (2015). Repetition in John Irving's novel a *Widow for One Year*. A corpus stylistic approach to literary translation. *International Journal of Corpus Linguistics*, 20(3): 355-377.

Čermáková, A. (2018). Translating children's literature: Some insights from corpus stylistics. *Ilha Desterro*, 71(1): 117-133.

Čermáková, A., & Fárová, L. (2010). Keywords in Harry Potter and their Czech and Finnish translation equivalents. In F. Čermák, P. Corness, & A. Klégr (Eds.), *InterCorp: Exploring a Multilingual Corpus* (pp. 177-188). Praha: NLN.

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 Wolfganf Teubert*. Amsterdam/Philadelphia: John Benjamins.
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.

Is death the great equalizer? A study of news accounts of women and men as murder victims

Monika Mondor (Gothenburg University) & Joe Trotta (Gothenburg University)
monika.mondor@sprak.gu.se; joe.trotta@sprak.gu.se

In this corpus-based study, we examine gender-based differences in how fatalities due to violent crimes are reported. Our primary research question is whether there are, encoded in the language used in newspaper texts, implicit biases in how such deaths are reported and whether these biases correlate to whether the victim is a man or a woman. Consider, for instance, the following two examples from newspapers (taken from the Corpus of Contemporary English or 'COCA'):

1. *The Lake County Coroner's Office has identified a woman found dead in her home […], and the […] Sheriff's Office announced charges against her husband of 18 years.*
2. *This is the worm farm co-owned by David Riess, who was found murdered on March 23.*

Wordings like those in the sentences above (*found dead* vs. *found murdered*) frame the events differently; our research analyzes how such framings are implemented, what variables (e.g., linguistic, contextual, situational, ideological) are involved, and what the possible reasons are for sex-based bias in news stories of this nature.

There is a considerable body of research within in academic fields such as media studies, journalism, women's studies, psychology, among others, on the ways in which violence against women is portrayed in the media (see, for example, Bullock, 2007; Ferraro, 2019; Jewkes, 2002; Taylor 2009). Such research provides a valuable point of departure for the present work, though these previous studies, because of their discipline-specific research goals, generally do not provide a sufficiently granular linguistic account of the data. Related research with a stricter linguistic focus tends to take on broader topics such as domestic violence in general (cf. Braber, 2015) and/or are limited to a small selection of texts/situations (cf. Buchner et al, 2021). Our study fills a gap in the research as it is distinctly linguistic in perspective, based on a systematic examination of corpus material, and explores accounts of fatalities of both male and female victims.

Using COCA and the British National Corpus (BNC), we investigate how the events are packaged linguistically; among other things we consider what synonyms, euphemisms or paraphrases are used to refer to the fatality (such as *murdered, slain, killed, found dead, found deceased*); whether there are explicit indications of a crime (i.e., being 'found murdered' indicates that one is a victim of a violent act, whereas being 'found dead' does not); and how, particularly in the case of female victims, discursive strategies are used that potentially 'mitigate' the crime (e.g. information about the decedent may include drug addiction and sex work).

Supported by the data from the corpus study, we interpret the findings through the lens of Critical Discourse Analysis to show how gender inequalities are constructed and signaled in news coverage of murder victims.

References

Braber, N. (2015). Representation of domestic violence in two British newspapers, The Guardian and The Sun, 2009-2011. *English Language Research Journal* (1), 86-104.

Buchner, V., Hamm, S., Medenica, B., & L Molendijk, M. (2021). *Linguistic analysis of online domestic violence testimonies in the context of COVID-19*. Advance. Preprint.
Bullock, Cathy Ferrand (2007). Framing domestic violence fatalities: coverage by Utah newspapers, *Women's Studies in Communication*, 30, 34-63.
Jewkes, Yvonne (2004). *Media and Crime*. London: Sage Publications.
Ferraro, F. R. (2019). Males tend to die, females tend to pass away. *Death Studies*, 43(10), 665–667.
Taylor, R. (2009). Slain and slandered: A content analysis of the portrayal of femicide in crime news. *Homicide Studies*, 13(1), 21–49.

Do formulaic sequences mask proficiency? Considering evidence from a large learner corpus

Akira Murakami, a.murakami@bham.ac.uk
Ute Römer (Georgia State University), uroemer@gsu.edu
Marije MichelDora Alexopoulou

Our paper seeks to explore the relationship between formulaic sequences (FSs) and second language (L2) proficiency, building on earlier work by Myles (2012). We define FSs as frequently-occurring combinations of words which may allow for internal variation and which carry stable meanings (e.g., 'make a decision', 'make a ADJ decision'). Earlier work suggests that the use of FSs positively affects fluency and accuracy, and that L2 learners benefit from an implicit or explicit focus on FS during instruction (Boers et al., 2006, Wray, 2018). Yet, when a learner's language use contains a large number of FSs, this might also create a challenge for language teachers and testers: when trying to assess the proficiency level, the structures underlying FSs may or may not reflect a student's level of lexicogrammatical competence. For example, a learner who uses FSs such as 'how do you do' or 'why don't you V' may not yet have mastered wh-question formation with verb inversion. This inspired us to explore whether (and if so in what ways) FSs mask the proficiency of L2 learners.

To address this question, we extracted data on a selection of high-frequency FSs from a 24-million word subset of the Cleaned Subcorpus of the EF-Cambridge Open Language Database (EFCAMDAT; Shatz, 2020) of learner writing. We started from lists of frequent n-grams of various lengths and grouped files in EFCAMDAT by learner proficiency, ranging from low beginner to advanced (CEFR levels A1 to C1). Two of the FSs selected for our analysis, 'why don't you V' and 'I think you should V,' first appear in A1 learner responses to a prompt in unit 21 (of 128) on 'Giving suggestions about clothing'. Despite their frequent early occurrence in the corpus, these two FSs build on fairly complex morphosyntactic structures (e.g., subject-verb inversion, *do*-support, negation). We studied the use of these and other FSs across EFCAMDAT levels with particular attention paid to their variability and accuracy of use. For each slot in a FS we measured the lexical diversity (MATTR, MTLD) and predictability (normalized entropy) at each proficiency level.

Results point to an overall lack of productivity in the use of the focus FSs at the A1 and A2 levels, as well as low accuracy when Al/A2 learners move away from the initial fixed sequences, resulting in erroneous or unidiomatic uses such as 'why don't you joining' and 'I suggest you should V'. For variable slots in the FSs, we found an increase in lexical diversity and a decrease in predictability as learners move from beginner to intermediate and advanced levels. We also observed that 'I think you should V' shows productivity earlier than 'why don't you V,' suggesting that learners need longer to acquire the more complex morphosyntax of this FS. In line with earlier work, we argue that FSs play an important role in early L2 development and instruction. Our findings suggest that assessment of language containing many FSs benefits from taking productivity and variability of their underlying structures into account.

References

Boers, F., Eyckmans, J., Kappel, J., Stengers, H., & Demecheleer, M. (2006). Formulaic sequences and perceived oral proficiency: Putting a lexical approach to the test. *Language Teaching Research*, *10*(3), 245-261.

Myles, F. (2012). Complexity, accuracy and fluency. The role played by formulaic sequences in early interlanguage development, (pp. 71-93) In Housen, A., Kuiken, F. & Vedder, I. (Eds.). *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA*.

Shatz, I. (2020). Refining and modifying the EFCAMDAT: Lessons from creating a new corpus from an existing large-scale English learner language database. *International Journal of Learner Corpus Research, 6*(2), 220-236. https://doi.org/10.1075/ijlcr.20009.sha.

Wray, A. (2018). Concluding question: Why don't second language learners more proactively target formulaic sequences?, (pp. 248-269). In Siyanova-Chanturia, A., & Pellicer-Sanchez, A. (Eds.). *Understanding formulaic language: A second language acquisition perspective*. Routledge.

Not that you asked, but here it is: a formal and functional taxonomy of *not-that* clauses in Present-day American English

Ozan Mustafa (University of Graz)
ozan.mustafa@uni-graz.at

In the past two decades there has been an increasing interest in various types of insubordination, elliptical constructions, and conventionalized fragments (e.g. Evans 2007; Goldberg & Herbst 2021). So far, *not-that* clauses as exemplified in (1), (2), and (3) have received very little attention. According to Schmid (2011, 2013) and Delahunty (2006), they are used as a means to avoid or clarify potential misunderstandings or, more generally, to reject a potential inference. However, the *not-that* clause appears to exhibit a range of further functions: for instance, retrospective, co-text dependent uses as in (1), prospective (contrastive) rhetorical uses as in (2), or hedging functions as in (3).

(1) *I have to say you're one of the best applicants I've ever interviewed.* ***Not that*** *it's a surprise.* (COCA)
(2) # ***Not that*** *you asked, but he's in surgery.* (COCA)
(3) *He didn't.* ***Not that*** *I know of.* (COCA)

This paper develops a taxonomy based on formal criteria and discourse-pragmatic functions. More specifically, it addresses the following questions: (i) how may this heterogeneous group be subclassified, (ii) how can they be best syntactically analyzed (i.e. ellipsis, insubordination, fragment), and (iii) how are they used in discourse (e.g. (1)-(3) above).

To answer these questions, the study also compares *not-that* clauses with their 'complete' cognates (i.e. 'inferentials'; see e.g. Delahunty 1990, 1995, 2001; Declerck 1992). The retrieved data suggests that these clauses differ with respect to their ability to be reconstructed as 'complete' cognates and in the nature of their relationship to the co(n)text or, more precisely, whether the *not-that* clause refers to the co-text like in (1) or discourse-context as in (2). Thus, whereas (1) may be rephrased as an inferential (4a), reconstructions such as in (4b-c) appear to be ill-formed. In addition, only the *not-that* clause with a hedging function may alternatively occur in a positive form, as illustrated in (5) (see also Hoeksama 2017).

(4) a. ***It's not that*** *it's a surprise.*
b. *?***It's not that** *you asked.*
c. ***It's not that** *I know of.*
(5) *He didn't,* ***that*** *I know of*.

The project uses the *Corpus of Contemporary American English* (COCA) to determine the morphosyntactic and discourse-functional features of the constructions in question. Due to their high frequencies in the COCA, a random sample of one thousand tokens each is taken and coded for their formal and functional properties. The data analysis will shed light on the differences and similarities between these constructions, as well as the varying degree of intersubstitutability. Thus, the paper aims to clarify, among others, whether the *not-that* clause needs to be considered to be an elliptical construct created online in the working memory or to be an entrenched (independent) construction in its own right. It is also argued that these constructions can be best accounted for in terms of

Construction Grammar, which models the relationships between different members of a constructional family (e.g. Diessel 2019; Van de Velde 2014; Goldberg 1995, 2006).

References

Davies, Mark. 2008-. *The Corpus of Contemporary American English* (COCA). Available online at www.english-corpora.org/coca/.

Declerck, Renaat. 1992. "The inferential it is that-construction and its congeners". *Lingua* 87: 203-230.

Delahunty, Gerald P. 1990. "Inferentials: The story of a forgotten evidential*". Kansas Working Papers in Linguistics* 15: 1-28.

Delahunty, Gerald P. 1995. "The inferential construction". *Pragmatics 5:* 341-364*.*

Delahunty, Gerald P. 2001. "Discourse functions of inferential sentences". *Linguistics* 39: 517-545*.*

Delahunty, Gerald P. 2006. "A relevance theoretic analysis of *not that sentences*: *'Not that there is anything wrong with that'*". *Pragmatics* 16(2/3): 213-245.

Diessel, Holger. 2019. *The grammar network: how linguistic structure is shaped by language use*. Cambridge: Cambridge University Press.

Evans, Nicholas. 2007. "Insubordination and its uses". In Nicolaeva, Irina (ed.). *Finiteness: Theoretical and Empirical Foundations*. Oxford: Oxford University Press, 366–431.

Goldberg, Adele E. 1995. *Constructions: A Construction Grammar approach to argument structure.* Chicago: The University of Chicago Press.

Goldberg, Adele E. 2006. *Constructions at work: the nature of generalization in language*. Oxford: Oxford University Press.

Goldberg, Adele E.; Herbst, Thomas. 2021. "The *nice-of-you* construction and its fragments". *Linguistics* 59(1), 285-318.

Hoeksema, Jack. 2017. "*Not that I know of*: a polarity-sensitive construction". *Linguistics* 55(6): 1281-1310.

Schmid, Hans-Jörg. 2011. "Tracing paths of conventionalization from the Bible to the BNC: a concise corpus-based history of the *not* that construction". In Renate Bauer & Ulrike Krischke (eds.). *More than words. English lexicography and lexicology past and present. Essays presented to Hans Sauer on the occasion of his 65th birthday, part I*. Frankfurt: Peter Lang, 299-316.

Schmid, Hans-Jörg. 2013. "Is usage more than usage after all? The case of English *not that*". *Linguistics* 51(1): 75-116.

Van de Velde, Freek. 2014. "Degeneracy: The maintenance of constructional networks". In Boogaart, Ronny; Colleman, Timothy; Rutten, Gijsbert (eds.). *Extending the Scope of Construction Grammar.* Berlin: De Gruyter Mouton, 141-180.

Concerns about cancer immunotherapy in online forum posts: A corpus-based discourse analysis

Hoa Ninh

This paper focuses on how online health community members linguistically frame their concerns about a relatively novel type of cancer treatment called cancer immunotherapy. Patients increasingly factor in the information they find on the Internet when making treatment decisions (Hardey 1999). Online health communities or fora are potential sites for people with health concerns to seek informational and emotional support or disclose personal problems that would be difficult to share in face-to-face interactions (White and Dorman 2001; Albrecht and Goldsmith 2003; Suler 2004; Prestin and Chou 2014; Hunt and Harvey 2015; Demjén 2016). As such, a growing body of linguistic research has examined how users of online health fora discursively encode their experiences (e.g. Stommel and Lamerichs 2014; Hunt and Koteyko 2015; Koteyko and Hunt 2018; Hunt and Brookes 2020). Many linguistic studies have explored online fora for cancer patients and carers (e.g. Sillence 2010,2013; Demmen et al. 2015; Demjén 2016; Potts and Semino 2017; Semino et al. 2017; Semino et al. 2018); however, among these studies, the topic of cancer treatment has been relatively under-researched. This is notable given that for cancer patients, the Internet is a commonly used source of information on the latest treatment protocols (Dickerson et al. 2006), and the experiences presented in online cancer communities play a role in patients' decision-making and coping processes regarding treatments and their side effects (Dickerson et al. 2006). Studies on patients' concerns about treatments have been conducted mostly in health and medical fields, focusing on *what* is said, rather than *how* it is said, about long-established forms of treatment (e.g. Beusterien et al. 2013; Freedman et al. 2016).

The present study focuses on immunotherapy – an umbrella term for an emerging group of cancer treatments targeting the immune cells – as it has received much attention from medical researchers (Couzin-Frankel 2013) as well as the media (Madden 2018) in recent years and received a Nobel Prize in 2018. My research question is: *How are concerns about cancer immunotherapy linguistically constructed in online forum posts?* Using a corpus-based approach to discourse analysis (Baker 2006), I examined the top collocates of the term *immunotherapy/ies* by conducting a qualitative analysis of the concordance lines of these collocates. The collocates were generated using a span of five words to the left and right of the search word. For the collocational measure, the cubed version of Mutual Information, or MI3, was selected as it has proved appropriate for forum data (Hunt and Brookes 2020). The posts were collected in October 2021 from an online community managed by a large research-focused cancer charity in the UK. The final dataset includes 1212 posts (over 252,000 words) dating back to July 2010, each containing at least one occurrence of the term *immunotherapy/ies* or its variants. The analysis reveals salient discussion points in these forum posts and how a range of problems or sources of uncertainty can be observed through contributors' narratives and requests for support.

References

Albrecht, T. L. and Goldsmith, D. J. 2003. Social Support, Social Networks, and Health. In: Thompson, T.L., Dorsey, A., Miller, K.I. and Parrott, R. eds. *Handbook of Health Communication*. Mahwah, New Jersey: Lawrence Erlbaum Associates, pp. 263-284.

Baker, P. 2006. *Using Corpora in Discourse Analysis*. London: A&C Black.

Beusterien, K., Tsay, S., Gholizadeh, S. and Su, Y. 2013. Real-world Experience with Colorectal Cancer Chemotherapies: Patient Web Forum Analysis. *Ecancermedicalscience* 7, pp. 361-361. doi: 10.3332/ecancer.2013.361

Couzin-Frankel, J. 2013. Cancer Immunotherapy. *Science* 342(6165), p. 1432. doi: 10.1126/science.342.6165.1432

Demjén, Z. 2016. Laughing at Cancer: Humour, Empowerment, Solidarity and Coping Online. *Journal of Pragmatics* 101, pp. 18-30. doi: 10.1016/j.pragma.2016.05.010

Demmen, J., Semino, E., Demjén, Z., Koller, V., Hardie, A., Rayson, P. and Payne, S. 2015. A Computer-assisted Study of the Use of Violence Metaphors for Cancer and End of Life by Patients, Family Carers and Health Professionals. *International Journal of Corpus Linguistics* 20(2), pp. 205-231.

Dickerson, S. S., Boehmke, M., Ogle, C. and Brown, J. K. 2006. Seeking and Managing Hope: Patients' Experiences Using the Internet for Cancer Care. *Oncology Nursing Forum* 33(1), pp. e8-e17. doi: 10.1188/06.ONF.E8-E17

Freedman, R. A., Viswanath, K., Vaz-Luis, I. and Keating, N. L. 2016. Learning from Social Media: Utilizing Advanced Data Extraction Techniques to Understand Barriers to Breast Cancer Treatment. *Breast Cancer Research and Treatment* 158(2), pp. 395-405. doi: 10.1007/s10549-016-3872-2

Hardey, M. 1999. Doctor in the House: The Internet as a Source of Lay Health Knowledge and the Challenge to Expertise. *Sociology of Health & Illness* 21(6), pp. 820-835.

Hunt, D. and Brookes, G. 2020. *Corpus, Discourse and Mental Health*. Bloomsbury Publishing.

Hunt, D. and Harvey, K. 2015. Health Communication and Corpus Linguistics: Using Corpus Tools to Analyse Eating Disorder Discourse Online. In: Baker, P. and McEnery, T. eds. *Corpora and Discourse Studies: Integrating Discourse and Corpora*. Hampshire: Palgrave Macmillan, pp. 134-154.

Hunt, D. and Koteyko, N. 2015. 'What was your blood sugar reading this morning?' Representing diabetes self-management on Facebook. *Discourse & Society* 26(4), pp. 445-463. doi: 10.1177/0957926515576631

Koteyko, N. and Hunt, D. 2018. Special Issue: Discourse Analysis Perspectives on Online Health Communication. *Discourse, Context & Media* 25, pp. 1-4. doi: 10.1016/j.dcm.2018.08.002

Madden, D. L. 2018. From a Patient Advocate s Perspective: Does Cancer Immunotherapy Represent a Paradigm Shift? *Current Oncology Reports* 20, pp. 1-7.

Potts, A. and Semino, E. 2017. Healthcare Professionals' Online Use of Violence Metaphors for Care at the End of Life in the US: A Corpus-based Comparison with the UK. *Corpora* 12(1), pp. 55-84. doi: 10.3366/cor.2017.0109

Prestin, A. and Chou, W.-y. S. 2014. Web 2.0 and the Changing Health Communication Environment. In: Hamilton, H.E. and Chou, W.S. eds. *The Routledge Handbook of Language and Health Communication*. Routledge, pp. 184-197.

Semino, E., Demjén, Z. and Demmen, J. 2018. An Integrated Approach to Metaphor and Framing in Cognition, Discourse, and Practice, with an Application to Metaphors for Cancer. *Applied Linguistics* 39(5), pp. 625-645. doi: 10.1093/applin/amw028

Semino, E., Demjén, Z., Demmen, J., Koller, V., Payne, S., Hardie, A. and Rayson, P. 2017. The Online Use of Violence and Journey Metaphors by Patients with Cancer, as Compared with Health Professionals: A Mixed Methods Study. *BMJ Supportive & Palliative Care* 7(1), p. 60. doi: 10.1136/bmjspcare-2014-000785

Sillence, E. 2010. Seeking out Very Like-minded Others: Exploring Trust and Advice Issues in an Online Health Support Group. *International Journal of Web Based Communities* 6(4), pp. 376-394. doi: 10.1504/IJWBC.2010.035840

Sillence, E. 2013. Giving and Receiving Peer Advice in an Online Breast Cancer Support Group. *CyberPsychology, Behavior & Social Networking* 16(6), pp. 480-485. doi: 10.1089/cyber.2013.1512

Stommel, W. and Lamerichs, J. 2014. Interaction in Online Support Groups: Advice and beyond. In: Hamilton, H.E. and Chou, W.S. eds. *The Routledge Handbook of Language and Health Communication*. Routledge, pp. 198-211.

Suler, J. 2004. The Online Disinhibition Effect. *CyberPsychology & Behavior* 7(3), pp. 321-326. doi: 10.1089/1094931041291295

White, M. and Dorman, S. M. 2001. Receiving Social Support Online: Implications for Health Education. *Health Education Research* 16(6), pp. 693-707. doi: 10.1093/her/16.6.693

Investigating Sentiments on Covid-19 in Tweets

Niklas Nitsch (TU Dortmund University) & Patricia Ronan (TU Dortmund University)
niklas.nitsch@tu-dortmund.de, patricia.ronan@tu-dortmund.de

Many corpora, both smaller, highly specialised and also extensive, large-scale ones, have been built using data from different social media platforms (e.g. Petrovic *et al*. 2010; Herdagdelen, 2013). Using social media data has the benefit of providing unfiltered user content, which is particularly useful for analysing opinions and sentiments. These corpora thus provide a good basis for sentiment detection (e.g. Kumar & Sebastian, 2012) and allow gathering insights into a user group's general stance on societally important topics.

With the persisting Covid-19 pandemic, opinions on measures implemented to contain the pandemic are shared and promoted on social media and can then influence opinions and sentiments of other users (Kaligotla *et al*. 2016). Analysing such shared sentiments amongst different individuals towards the current global crisis can help us to understand the underlying societal processes better (cf. Nagy & Stamberger, 2012). For this, social media in general, and tweets in particular, have previously been used beneficially (e.g. Barbosa & Feng, 2010). The current study aims to investigate sentiments conveyed in tweets towards different opinion target groups relevant to the Covid-19 pandemic.

Twitter is chosen as a data source as it widely accessible and far reaching and offers any user the possibility to voice their opinion publicly, and it provides easily accessible data. The present study uses a small purpose-built corpus of just over 1800 English-language tweets from late Covid periods, starting around April 2022. They have been collected using Twitter's back-end API. The collection is based on two factors: 1) tweets must include specific Covid-19-related hashtags, for example *#COVID19*, and 2) have been posted after the initial data collection stage, from April 2022 onwards. In this study, the harvested tweets are first manually categorized according to the opinion target of the sentiment. Manual categorisation follows a rigorous multi-step process, which is a further reason for the small corpus size. Then, using the R package "sentimentr" (Rinker, 2021) and a sentiment dictionary by Mohammad & Turney (2013), as well as R code by Schweinberger (2022), we assign sentiment categories to words in the tweets.

The analysis shows that significantly differing sentiment categories towards eight relevant opinion target groups can be observed, such as "Government & Politicians", "Science & Healthcare" or "Media", among others. Frequency of opinions towards these assigned targets differs substantially, as well as frequencies of different sentiments. Negatively polarised sentiments, most prominently 'fear', have a large share. However, we can also see frequent occurrences of positively polarised sentiments like 'trust' or 'anticipation' throughout opinion target groups covering, for example, personal communication. Through this small-scale approach, relevant differences can already be detected, for example: negative polarisation of 5 out of 8 opinion targets, prevalence of 'fear' for negative and 'trust' for positive sentiments, and the dominance of the word 'pandemic' as the major influence of negative polarisation. Finally, this study highlights the effectiveness of manual classification in combination with automatic sentiment tagging to achieve more insightful and interpretable results.

References

Barbosa, L. & Feng, J. 2010. Robust sentiment detection on Twitter from biased and noisy data. In: *Proceedings of the 23rd International Conference on Computational Linguistics, ICCL*, 36-44.

Herdagdelen, A. 2013. Twitter n-gram corpus with demographic metadata. *Language Resources and Evaluation*, 47, 1127-47. doi: 10.1007/s10579-013-9227-2

Kaligotla, C., Yucesan, E. & Chick, S. E. 2016. The impact of broadcasting on the spread of opinions in social media conversations. In: T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, & S. E. Chick (eds.), *Proceedings of the 2016 Winter Simulation Conference*, 3476-87. doi: 10.1109/WSC.2016.7822377

Kumar, A. & Sebastian, T. M. 2012. Sentiment Analysis on Twitter. *International Journal of Computer Science Issues*, 9.3, 372-8.

Mohammad, S. M. & Turney, P. D. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29.3, 436-65. doi: 10.48550/arXiv.1308.6297

Nagy, A. & Stamberger, J. A. 2012. Crowd sentiment detection during disasters and crises. In L. J. M. Rothkrantz, J. Ristvej, Z. Franco (eds.), *9th Proceedings of the International Conference on Information Systems for Crisis Response and Management*.

Petrovic, S., Osborne, M. & Lavrenko, V. 2010. The Edinburgh Twitter Corpus. In: B. Hachey & M. Osborne (eds.), *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*. Association of Computational Linguistics, 25-6.

Rinker, T. 2021. *'sentimentr': Calculate Text Polarity Sentiment*. [R package]. Available at: https://cran.r-project.org/web/packages/sentimentr/index.html. Accessed on: 13/05/2022.

Schweinberger, M. 2022. Sentiment Analysis in R. *LADAL*. The University of Queensland. Available at: https://slcladal.github.io/sentiment.html. Accessed on: 16/05/2022.

A Closer Look at GET in African Postcolonial Englishes

Temitayo Olatoye (University of Eastern Finland)

Get is a highly polysemous verb in Present Day English (PDE). In the perfect construction, gotten as the past participle of *get* has been described as a morphological Americanism in PDE (Anderwald 2021). The use of *got* vs. *gotten* has been associated with stative vs. dynamic possession in American English (AmE), whereas British English (BrE) primarily employs *got* in both cases. Unlike the perfect construction, the passive construction is relatively infrequent in PDE. In recent times, research has established a decrease in prototypical *be*-passives and an increase in informal *get*-passives in both AmE and BrE, with AmE in the lead (Leech et al. 2009, Collins & Yao 2013). This increase in the usage of informal variants has been ascribed to a growing degree of colloquialization in PDE (Mair 1997), and recent research suggests that AmE is at the vanguard of ongoing changes towards more colloquial grammatical features.

At the end of the twentieth century, a growing American influence was reported in several African postcolonial Englishes (Awonusi 1994, Shoba et al. 2013). Although a few researchers have examined AmE influence in lexis and phonology, grammatical structures have been less researched using corpus data. Against this background, the present study investigates the use of *get* in passive, modal and possessive constructions as in: he got rewarded, he has got to be rewarded, and he has got(ten) his reward in two West African Englishes, two East African Englishes, their shared historical input variety (BrE) and the twentieth century most influential variety (AmE).

Using a corpus-based approach, spoken and written data from the Ghanaian, Nigerian, Kenyan, Tanzanian, British and American components of the International Corpus of English (ICE), and the Santa Barbara corpus of spoken AmE are analysed quantitatively and qualitatively in order to answer the following research question: To what extent (and perhaps why) do usage patterns associated with the get-passive, quasi-modal and possessive constructions differ in these varieties of English?

Results are expected to confirm a growing influence of AmE in norm-developing Englishes, provide evidence for a regional contrast in the use of these three *get* constructions and enhance our understanding of African postcolonial Englishes.

References

Anderwald, Lieselotte. 2021. The Complex History of Have Gotten in American English. *American Speech*, 1-46.

Awonusi, Victor O. 1994. The Americanization of Nigerian English. *World Englishes*, 13: 75–82.

Collins, Peter & Xinyue Yao. 2013. Colloquial features in World Englishes*. International Journal of Corpus Linguistics.* 18(4). 479–505.

Leech, Geoffrey, Marianne Hundt, Christian Mair & Nicholas Smith. 2009. *Change in contemporary English: A grammatical study*. Cambridge: Cambridge University Press.

Mair, Christian. 1997. Parallel Corpora: A Real-Time Approach to the Study of Language Change in Progress. In Magnus Ljung, ed. *Corpus-based Studies in English*. Amsterdam: Rodopi, 195–209.

Shoba, Jo, Dako, Kari and Orfson-Offei. 2013. 'Locally acquired foreign accent' (LAFA) in contemporary Ghana. *World Englishes*, 32: 230–242.

Spelling variation in inner-circle Englishes

Marta Pacheco-Franco (Universidad de Malaga)
martapacheco@uma.es

Despite the variation found in Present-day English, its spelling system continues to be described as showing polarisation between British English and American English, both in academic (Gramley et al. 2021; Baker 2017; Peters 2004, 2007; Huddleston and Pullum 2003) and non-academic sources (see among others English Club n.d.; Mondly n.d.; Confused Words n.d.). Underlying this dichotomy is the assumption that the remaining World Englishes will follow either one written standard or the other, which also includes inner-circle varieties. Although these have mostly reached endonormative stabilisation along Schneider's (2007) cycle, external norms continue to govern the spelling of Australian, Canadian, Irish and New Zealand Englishes. Their shared historical pasts and their persisting cultural links with the United Kingdom suggests that these varieties ought to follow BrE spelling norms, with anecdotal occurrences of AmE forms (Melchers and Shaw 2011; Trudgill and Hannah 2008). However, recent studies on the overall Americanisation of English seem to raise some questions on the reality of orthography today. How do spelling variants distribute in the inner circle varieties? What does their distributional patterns signify?

The present paper aims to answer these questions by means of a corpus-based investigation that will analyse the spelling of the inner-circle varieties. In order to do so, the study will draw on the distribution of the three spelling variants that best represent the tensions between British and American English orthographic norms: namely, *-our*/*-or* as in *colour*, *-re*/*-er* as in *theatre* and *-isation*/*-ization* as in *realisation*. The source material comes from the *Global Web-based English* or *GloWbE* corpus (Davies 2013), which contains 1.9 billion words of text produced on the internet for twenty different varieties of English. These features make the corpus an excellent resource for the diatopic analysis of English language use online. The data will be gathered from the *Global Web-based English* or *GloWbE* corpus in a three-staged process, which involves (1) retrieving the complete lists of occurrences with either spelling variant, (2) selecting the input for the study and (3) gathering the quantitative data, a total of 1,980,565 tokens. The analysis so far has shown that, although the British variants remain dominant in most varieties, the American forms seem to be playing an important role in Canadian and Australian Englishes. On the one hand, Canadian English spelling is on the way to becoming an amalgam of British and American forms, which seems to illustrate the country's own liminality. On the other hand, Australian English spelling remains conservative. However, some American forms have been assimilated into the variety so as to convey some distinct meanings. These results, though tentative, shed some light on the impact of Americanisation not only on orthography, but on language as a whole.

References

Baker, Paul. 2017. *American and British English*. Cambridge: Cambridge University Press.
Confused Words. n.d. "British vs American Spelling Differences ESL Learners Should Know." [Accessed April 26, 2021].
Davies, Mark. 2013. "Corpus of Global Web-Based English (GloWbE)." [Accessed January 28, 2022].
English Club. n.d. "British and American Spelling." [Accessed April 26, 2021].

Gramley, Stephan, Vivian Gramley and Kurt-Michael Pätzold. 2021. *A Survey of Modern English*. London: Routledge.
Huddleston, Rodney and Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.
Melchers, Gunnel and Philip Shaw. 2011. *World Englishes*. 2nd ed. London: Hodder Education.
Mondly. n.d. "British English vs American English Differences: Spelling, Pronunciation and More." [Accessed April 26, 2021].
Peters, Pam. 2004. *The Cambridge Guide to English Usage*. Cambridge: Cambridge University Press.
—. 2007. *The Cambridge Guide to Australian English Usage*. Cambridge: Cambridge University Press.
Schneider, Edgar W. 2007. *Postcolonial English: Varieties around the World*. Cambridge: Cambridge University Press.
Trudgill, Peter and Jean Hannah. 2008. *International English: A Guide to Varieties of Standard English*. 5th ed. London; New York: E. Arnold.

Direct object definiteness and verb meaning: A corpus-based investigation

Florent Perek (University of Birmingham) & Lotte Sommerer (University of Freiburg)
florent.perek@gmail.com; lotte.sommerer@anglistik.uni-freiburg.de

Research on English and other languages typically makes two basic assumptions about the grammar of verbs and nouns. First, verbs are taken to determine the morphosyntactic category of the constituents they occur with (e.g. Tesnière 1959). Verb meanings in particular have been found to play a major role in argument realization (Levin 1993, Levin & Rappaport Hovav 2005). Second, (in)definiteness marking is seen as a discourse-pragmatic feature which (in English) is overtly and obligatorily coded by determiners, in particular the articles *the* and *a/an*, and signals to the hearer whether a referential NP should be familiar to them and/or has been talked about before (e.g. Hawkins 1978, Lyons 1999). Argument realization and definiteness marking are typically seen as separate and unrelated grammatical phenomena.

This paper takes a new look at these two areas of grammar and uses corpus data to investigate the relation between verbs and the definiteness of one of their arguments, specifically the direct object. On the basis of a large corpus of 3.4 million direct object NPs extracted from the British National Corpus (XML Edition) by means of a dependency parser (Chen & Manning 2014), we find the relative frequency of definite vs. indefinite direct objects to vary widely according to the verb. This variation can be related to the meaning of the verb, in that verbs with a similar meaning, as measured by a distributional semantic model (Lenci 2018), tend to occur to a similar extent with (in)definite direct objects, and the preference of verbs for (in)definite objects can often be explained by some of their semantic properties. For example, verbs like *produce* or *need* are much more likely to combine with an indefinite NP than with a definite NP due to the fact that one produces or needs something that one does not yet possess and hence is most likely unfamiliar with (ex. 1). In contrast, a verb like *explain* or *forget* is highly likely to collocate with a definite NP because one can only explain or forget what one is already aware of, and thus is familiar and specific (ex. 2).

(1) The first round of the 1981 American Open **produced <u>a surprise that should be recorded</u>** (BNC, HJG)
(2) The package **explains <u>the complexities</u>** of serial communications using on screen tutorials with animated sections (BNC, HAC)

Our data show that along with intertextual discourse reasons, the semantics of a particular verb seems to have an influence on the so-called 'definiteness profile' of the arguments it licenses. This suggests that argument realization and definiteness marking might not be as separate as it is usually assumed. From a usage-based point of view (e.g. Goldberg 2006, Perek 2015), we hypothesize that verbs project not only information about the morphosyntactic encoding of their arguments, but also expectations about their discourse status.

References

Chen, D. and Manning, C. (2014). A Fast and Accurate Dependency Parser Using Neural Networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 740-750. Doha, Qatar: ACL.

Lenci, A. (2018). Distributional Models of Word Meaning. *Annual Review of Linguistics* 4(1): 151-171.
Levin, B. (1993). *English Verb Classes and Alternations*. Chicago: University of Chicago Press.
Levin, B. & Rappaport Hovav, M. (2005). *Argument Realization*. Cambridge: Cambridge University Press.
Goldberg, A E. (2006). *Constructions at Work. The Nature of Generalization in Language*. Oxford: Oxford University Press.
Hawkins, J. A. (1978). *Definiteness and Indefiniteness. A Study in Reference and Grammaticality Prediction*. London: Croom Helm/Routledge.
Lyons, C. (1999). *Definiteness*. Cambridge: Cambridge University Press.
Perek, F. (2015). *Argument structure in usage-based construction grammar: Experimental and corpus-based perspectives*. Amsterdam: John Benjamins.
Tesnière, L. 1959. *Éléments de syntaxe structurale*. Paris: Klincksieck.

F0 range in L2 discourse: A corpus-based contrastive interlanguage analysis

Karin Puga (Justus Liebig University Giessen)
karin.puga@anglistik.uni-giessen.de

The concept of interlanguage, coined by Selinker in 1972, has attracted immense scholarly attention, particularly since Granger's (2015) Contrastive Interlanguage Analysis was published. The investigation of the interlanguage systems of learners from different L1 backgrounds has covered many linguistic areas, e.g. phraseology, syntax, and pragmatics. However, research into interlanguage prosody in general and the f0 range in particular has remained an exception. Overall, the fundamental frequency (f0) range in L2 speech is narrower than L1 English speech, irrespective of the learners' L1, speaking style, and speech function (e.g. Ramírez-Verdugo 2022; Gut 2009; Volín et al. 2015). Many scholars attribute their results to L1 influence, uncertainty, or a lack of confidence, and other explanations are rarely offered. L1 influence is postulated because learners often produce an f0 with intermediate values between their own L1 and those of native speakers, their f0 span deviating more greatly than their f0 level. The present study seeks to answer the following research questions:

1. Is f0 range always narrower in L2 speech?
2. Are there alternative interpretations of a deviating f0 range?

A mixed-methods approach and a multivariate analysis are adopted in the examination of L1 (n=90) and L2 data (n=135). The database consists of prosodically annotated versions of the Czech, German, and Spanish components of LINDSEI, alongside British (LOCNEC) and American English (NWSP & NSV) control corpora. Using an autosegmental-metrical approach (Beckman & Pierrehumbert 1986), the study investigates acoustic properties of the f0 range (level and span) of declarative utterances extracted from spontaneous speech (dialogic and monologic) on similar topics (personal narratives). Regression modeling was used to predict the f0 range of tune patterns by L1/L2 speaker groups and to investigate the effect of several (extra)linguistic factors, e.g. gender, L2 proficiency (B1-C2: based on post-hoc ratings by Huang et al. 2018), duration of stay abroad, speaking style, and intermediate phrase length in the f0 range.
The results show that, while learners approximate their targets for high-low tunes (a high pitch accent at the beginning of an intermediate phrase ending in a low tone) at the f0 level, they produce a significantly narrower f0 span for the same tunes (-0.7 to -1.9 semitones). Further significant tune-based differences in L2 speech are higher and wider high-ending tunes (1-2 semitones).

A combination of (extra)linguistic variables explains the results; for instance, female L2 speech deviates more than male L2 speech, and the longer the intermediate phrases, the higher and wider the f0. L1 influence cannot be ruled out as a factor determining the narrower f0 in high-low tunes (underhitting) and higher and wider f0 in low-high tunes (overhitting). However, L2 proficiency levels seem to be more revealing; all the learners manifested similar trends, but deviation from native speakers was more pronounced in the lower-proficiency learner group. Besides signaling insecurity, the extremely high f0 range produced in high-ending tunes in L2 speech was also found to fulfill a discourse management function to possibly compensate for weaknesses in intonational phrasing and to make cohesion between intonation units more explicit.

References

Beckman, M. E., & Pierrehumbert, J. (1986). Intonational structure in Japanese and English. *Phonology Yearbook*, 3(1), 255-309.

Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research*, 1(1), 7-24.

Gut, U. (2009). *Non-Native Speech: A Corpus-Based Analysis of Phonological and Phonetic Properties of L2 English and German*. Frankfurt: Peter Lang.

Huang, L.-F., Kubelec, S., Keng, N. & Hsu, L.-H. (2018). Evaluating CEFR rater performance through the analysis of spoken learner corpora. *Language Testing in Asia*, 8(14), 1-17.

Ramírez-Verdugo, M. D. (2022). *Intonation in L2 Discourse. Research Insights*. New York: Routledge Studies in Applied Linguistics.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics in Language Teaching*, 10(3), 209-231.

Volín, J., Poesová, K., & Weingartová, L. (2015). Speech melody properties in English, Czech and Czech English: Reference and interference. *Research in Language*, 13(1), 107-123.

A gender-based analysis of pragmatic markers in Sri Lankan English

Mahishi Ranaweera

Popular science often professes strong claims on differences between the speech patterns of men and women in spite of vigorous argument against such claims. For example, it is often presented that women tend to talk or use certain language features more than men (Tannen; 1992). There is also much empirical research attesting that speakers are acutely aware of what is gender appropriate behaviour in their speech (Holmes, 2014; Buzzanell & Meisenbach 2006; Sauntson, 2018). This is reflected in the choice of language structures they use (Lange & Leuckert,2020). In the use of discourse markers and pragmatic markers, research claim that females lead men (Macaulay, 2002). As an empirical contribution to this debate, the present corpus-based comparative study analyses the use of pragmatic markers (PMs) by both men and women who speak Standard Sri Lankan English (SSLE).

More specifically this paper considers how (and which) PMs are used by SSLE speaker and, whether there is a gender difference in their production. This study, part of a larger project, is based on a corpus of approximately 66,000 words consisting of conversations from 12 men and 12 women participating in talk show series in YouTube channels between 2016-2021. Each script in the corpus is approximately 2750 words with an average of 2200 words spoken by the interviewee. All speakers come from a similar occupational setting. However, they are from a mixed age group belonging to four different ethnic identities to capture the diverse social contexts of the SSLE speakers.

The frequency and the type of PMs, patterns of use, and their context were identified using AntConc. The functions of the PMs were assigned manually following the categories outlined in previous research (Aijmer, 2013; Beeching, 2016). The quantitative perspective was complimented by qualitative perspectives.

This preliminary analysis has identified 15 distinct PMs and 7 distinct co-occuring PMs (i.e. 'like you know', 'sort of you know', 'well I mean', 'kind of like/like kind of', 'you know I mean/ I mean you know', 'Okay like', and 'Right now'). The speakers used an average of 9 different PMs in a one script, with 'you know' being the most frequently used PM by all speakers across many functions, predominantly to fill pauses. There are some differences in the choice of other PMs; e.g. 'you see' is used more frequently by men and 'like' more frequently by women. However, overall, the PMs are remarkably more common among male than female speakers. This contrasts with the observations of previous research that women tend to cooperate and use hedging more than men as PMs are broadly used to add vagueness to maintain dialogue. In terms of function, the PMs show similarity across both groups suggesting that at least in this group of speakers, gender may not be the primary factor affecting the reason for the use of PMs. There is also the evidence of a nativized PM marker 'no' used to mark assertion or shared previous knowledge. This is mainly used in lieu of the Sinhala language PM 'ne' and further contributed to mark intonation similar to Sinhala language.

References

Aijmer, Karin. 2013. *Understanding pragmatic markers: A variational pragmatic approach*,
Edinburgh: Edinburgh University Press.

Beeching, Kate. 2016. *Pragmatic Markers in British English*. Cambridge University Press.

Holmes, Janet. 2014. Language and Gender in the Workplace. In Ehrlich, Meyerhoff and Holmes (eds.), The Handbook of Language, Gender, and Sexuality, Second Edition, 433-451. John Wiley and Sons, Ltd.

Buzzanell, Patrice M., & Rebecca Meisenbach. 2006. Gendered Performance and Communication in the Employment Interview. In Barrett and Davidson (eds.), Gender and Communication at Work, 19–37. Aldershot: Ashgate.

Lange, Claudia & Leuckert, Sven. 2021. Tag Questions and Gender in Indian English. In Bernaisch (ed.), Gender in World Englishes, 69-93. Cambridge: Cambridge University Press. doi:10.1017/9781108696739.004

Sauntson, Helen. 2018. Language, sexuality and education. Cambridge: Cambridge University Press.

Tannen, Deborah. 1992. You just don't understand: women and men in conversation. London: Virago.

Grinding to a halt? The spread of the progressive in recent spoken British English

Paula Rautionaho (University of Eastern Finland)
paula.rautionaho@uef.fi

The historical spread of the progressive form (BE + V*ing*) has been documented comprehensively in studies such as Leech et al. (2009), Kranich (2010), Smith (2005), and Smitterberg (2005) – the rise of the progressive during the 17th and 18th centuries was followed with a rapid increase in the 19th century, but the 20th century saw a gradual levelling. While most previous studies focus on written language, Smith (2005) shows that, in spoken data, the frequency of the progressive between the 1960s and 1990s remains stable, possibly indicating that the plateau of the S-curve had already been reached. The two versions of the *British National Corpus* now provide an opportunity to investigate how the frequency of the progressive fares between the early 1990s and the 2010s: is the spread of the progressive grinding to a halt after 400 years?

This study investigates the progressive form in the original demographic sample of BNC (BNC1994DS) and the sample release of the new version (BNC2014S), focusing on the frequency of the construction and the lemmata co-occurring with it. All present participles preceded by a form of the auxiliary BE, with up to 3 intervening words, were extracted from the corpora, and the relevance of the extracted tokens was manually checked (excluded instances include BE *going to*, adjectival and gerundial forms, and *to*-infinitives). In order to assess the distinctiveness of the individual lemmata occurring in the two datasets, a distinctive collexeme analysis (see e.g. Levshina 2015: 241-8) was performed in R. In addition to changes in the frequency of individual lemmata, the study focuses on a number of more or less fixed patterns involving the progressive (see Rohe 2018; also Römer 2005).

The results, based on a database of 66,595 progressives, indicate that the plateau has indeed been reached and the progressive is no longer increasing in frequency; BNC1994DS has 748.4 progressives per 100,000 words, whereas BNC2014S has 728.9 (this decrease is statistically significant at $p < 0.001$). The ten most frequent lemmata are the same in the two corpora (e.g. *going*, *doing*, and *coming*), but the distinctive collexeme analysis indicates either increasing or decreasing use of individual lemmata co-occurring with the progressive. Overall, the trends indicate that the frequency of dynamic (e.g. *going*, *coming*) and stance verbs (e.g. *standing*, *sitting*) has decreased in the timespan investigated, while stative (e.g. *being*, *loving*) and communication verbs (e.g. *saying*, *talking*) show increasing frequencies. The increasing number of fixed patterns involving the progressive form, such as *I BE just saying*, *I BE just thinking/wondering*, *I am/'m not saying*, and *I am/'m not being*, may indicate that the progressive is acquiring more robust pragmatic, i.e. non-aspectual, uses in recent spoken British English, despite the overall decrease in progressive frequency.

References

Kranich, Svenja. 2010. *Progressive in Modern English: A Corpus-based Study of Grammaticalization and Related Changes.* Amsterdam: Rodopi.

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

Levshina, Natalia. 2015. *How to do Linguistics with R. Data Exploration and Statistical Analysis.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

Rohe, Udo. 2018. The progressive in present-day spoken English. Real-time studies of its spread and functional diversification. PhD dissertation, University of Freiburg.
Römer, Ute. 2005. *Progressives, Patterns, Pedagogy: A Corpus-Driven Approach to English Progressive Forms, Functions, Contexts and Didactics.* Amsterdam/Philadelphia: John Benjamins Publishing Company.
Smith, Nicholas Ian. 2005. A Corpus-based investigation of recent change in the use of the progressive in British English. PhD dissertation, Lancaster University.
Smitterberg, Erik. 2005. *The Progressive in 19th-century English: A Process of Integration.* Amsterdam: Rodopi.

Another turn of the screw on the development of *-ity* and *-ness* across registers: Early and Late Modern English periods in focus

Paula Rodríguez-Puente(University of Oviedo)
rodriguezppaula@uniovi.es

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 and Late Modern English periods and across a wide range of registers distributed along a formal-informal and speech-written continuum. Despite the importance of register analysis in the development of languages (Biber & Gray 2013), few investigations had so far explored the interplay between suffix usage and register during these periods. Based on evidence from *A Representative Corpus of Historical English Registers* (ARCHER), Cowie (1998) measured the aggregation of new types (see Cowie & Dalton-Puffer 2002) of the two suffixes from 1650 to 1990, concluding that register is not determinant in their use. Seeking to verify Cowie's (1998) results, Rodríguez-Puente (2020, 2021) based her results on a larger sample of registers represented in a more modern version of ARCHER (3.2), *A Corpus of English Dialogues 1560-1760*, the *Penn-Helsinki Parsed Corpus of Early Modern English*, the *Penn Parsed Corpus of Modern British English*, the *Corpus of Historical English Law Reports, 1535-1999* and a sample of the *Old Bailey Corpus*. Measuring types and aggregation of new types, Rodríguez-Puente's (2020) results suggest that *-ity* gained ground on *-ness* between the sixteenth and eighteenth centuries. This change begins in formal written registers and spreads 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), which would arguably favor the use of the more learned and prestigious borrowed form (see also Rodríguez-Puente et al. 2022). Applying similar measures, Rodríguez-Puente (2021) notes a change in tendencies during the Late Modern English period. Whereas, during the eighteenth and early nineteenth centuries, *-ity* is still the predominant suffix, between the second half of the nineteenth and the early twentieth centuries, speech-related registers generally tend to turn to *-ness*, as if returning to a less elaborate and literate style, which may have been favored by the progressive democratization of language and the colloquialization of written registers (Hundt & Mair 1999; Hiltunen & Loureiro-Porto 2020).

Using Rodríguez-Puente's (2020, 2021) dataset, this paper seeks to further investigate the productivity of the two suffixes across registers, this time paying attention to the internal structure of the formations (see Säily et al. 2021). More precisely, I intend to analyze 1) whether the word was borrowed already bearing the suffix or whether the derivative was created in English; 2) the etymology of the base; and, 3) the part of speech of the base. I also seek to compare doublets (e.g. *denseness-density*) to ascertain whether, under two possible available options, the Romance suffix is preferred in formal, writing-based registers. Preliminary findings suggest that, in general, *-ness* is more versatile in that it combines both with native and borrowed bases belonging to different parts of speech, and produces more types derived within English. However, differences in the features of the two suffixes are notable across registers.

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, Douglas & Bethany Gray. 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.

Cowie, Claire & Christiane Dalton-Puffer. 2002. Diachronic word-formation and studying changes in productivity over time: Theoretical and methodological considerations. In Javier Díaz-Vera (ed.), *A Changing World of Words. Studies in English Historical Lexicography, Lexicology and Semantics*, 410-437. Amsterdam: Rodopi.

Hiltunen, Turo & Lucía Loureiro-Porto. 2020. Democratization of Englishes: Synchronic and diachronic approaches. *Language Sciences* 79: 101275. 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.

Rodríguez-Puente, Paula. 2021. Suffix competition across registers: On the development of *-ity* and *-ness* in the Late Modern English period. Paper presented at the ICAME42 Conference, Dortmund, 18-21 August 2021.

Rodríguez-Puente, Paula, Tanja Säily & Jukka Suomela. 2022. New methods for analysing diachronic suffix competition across registers: How *-ity* gained ground on *-ness* in Early Modern English. *International Journal of Corpus Linguistics*. https://doi.org/10.1075/ijcl.22014.rod

Säily, Tanja, Martin Hilpert & Jukka Suomela. 2021. New approaches to investigating change in derivational productivity. Paper presented at the ICAME42 Conference, Dortmund, 18-21 August 2021.

What's Normal in Conversations? Non-Canonical Interrogatives in LINDSEI-EST Corpus

Kärt Roomäe (University of Birmingham)
kxr177@student.bham.ac.uk

This presentation reports on the work on non-canonical interrogatives (e.g., rhetorical questions) in the Estonian subcorpus of the Louvain International Database of Spoken English Interlanguage (henceforth, LINDSEI-EST). The corpus is the first of its kind in Estonia and the data has been previously studied from the viewpoint of intensifiers in English as L2 (Klavan, Roomäe, and Savchenko 2020: 125-129). The research question of this study is: What non-canonical formal and/or functional features are present in LINDSEI-EST corpus? The hypothesis of this work-in-progress report is that the dynamicity of grammar (cf. Hopper 1987: 144) leads to irregularity in usage.

Conversations in LINDSEI-EST are intersubjective, and they are based on interviews that last for approximately 15 minutes, involving three tasks: a monologue on a chosen topic followed by spontaneous dialogue and picture description. The corpus is a work-in-progress; currently, 25 interviews have been recorded and transcribed, amounting to approximately 346 min. The interviewees are 3rd or 4th-year students of English Language and Literature BA program at the University of Tartu, Estonia. All are native speakers of Estonian. Besides the focus on the relationship of dynamicity and irregularity in usage, this study is also interested in the effect of the setting that is informal but also presents a relatively guided narrative. The spontaneity of the interviews decreases somewhat as a result. Corpus linguistics shows potential in studying the patterns in the dataset.

By doing a qualitative study on the 342 interrogative utterances in the interviews, non-canonical structures are expected. Among the 223 non-canonical interrogatives, irregular word order, repetition, ellipsis, like the alternative interrogative in *<A> <overlap /> did you spend there the whole year or or </A>* (Interview 20), but also functional characteristics such as the tag in *<A> (mm) but you are on your third year of BA studies right now right </A>* (Interview 10) highlight the complexity of conversations. To some extent, there is also Estonian L1 influence in terms of word order. As Dayal (2016: 268) points out, the non-canonical features can also be combined within a single interrogative utterance. Therefore, it might be the case that in addition to non-canonical word order, the speech act fulfills an additional function or has a distinct function altogether, unrelated to the speech act of asking for information. The speaker's purpose leads to the use of a more complex form (ibid.).

Non-canonical interrogative utterances in LINDSEI-EST also show the complementarity of discourse and syntax. My analysis considers interaction, therefore largely following the premises of Interactional Construction Grammar (e.g., Fischer 2015; Imo 2015), and adds to this new field of study. For functionalists, it is crucial to investigate what is happening beyond the level of the sentence, especially as so far, the discourse "has largely remained a waste-paper basket" in terms of constructionist frameworks (Östman 2005: 125).

References

Dayal, Veneeta. 2016. *Questions.* Oxford: Oxford University Press.

Fischer, Kerstin. 2015. Conversation, Construction Grammar, and cognition. *Language and Cognition*, 7, 563–588.

Hopper, Paul. 1987. *Emergent Grammar*, 139-157. In *Proceedings of the Thirteenth Annual Meeting of the Berkeley Linguistics Society.*

Imo, Wolfgang. 2015. Interactional Construction Grammar. *Linguistics Vanguard*, 1:1, 69–77.

Klavan, Jane; Roomäe, Kärt; and Savchenko, Denys. 2020. A corpus study of intensifiers in Estonian learners' spoken English. In Busse, Beatrix; Dumrukcic, Nina; and Möhlig-Falke, Ruth (ed). *Extended Book of Abstracts of The International Computer Archive of Modern and Medieval English annual conference ICAME 41,* 125–129. University and City Library (USB) Cologne: Cologne University Publication Server KUPS.

Östman, Jan-Ola. 2005. Construction Discourse: A prolegomenon. In Östman, Jan-Ola and Mirjam Fried (ed). *Construction Grammars: Cognitive grounding and theoretical extensions,* 121-144. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Word order in additional-language English spoken by multilinguals

Sylvi Rørvik (Inland Norway University Of Applied Sciences)
sylvi.rorvik@inn.no

This paper reports on a study of word order in the spoken English of seven Congolese refugees who have recently arrived in Norway, and are thus searching for "their new normal". English is likely to be these speakers' primary language of communication in Norway before they acquire a working command of Norwegian, so a description of their word order preferences in English is a useful starting point for future studies of potential cross-linguistic influence from their English production to their Norwegian production.

The seven participants share language backgrounds to the extent that they speak one or more Bantu languages, e.g. Swahili and/or Kinyabwisha, and they all speak English. Furthermore, they acquired all or most of their knowledge of English while living in Uganda. It was therefore hypothesized that their English production would generally comply with the syntax of Standard English, but that it would also display features are characteristic of Ugandan English and Bantu languages. Specifically as regards word order, previous research indicates that one may expect to find a greater frequency of left-dislocation than in Standard English (cf. e.g Mesthrie 1997; Schmied 2006, 2008; Makalela 2007; Nassenstein 2016; Ssempuuma 2017), and indeed it has been asserted that left-dislocation is a "feature characterizing most New Englishes" (Meierkord 2004: 128). Left-dislocation is infrequent in Norwegian, and transfer of this feature would therefore constitute a marked feature in the interlanguage of these seven speakers. The research question was thus: In the spoken English of Congolese refugees, is there evidence of left-dislocation which complies with what one might expect from English learned in Uganda?

The definition of left-dislocation employed roughly corresponds to that provided by Biber et al (1999: 138), and comprises an element (often a noun phrase) placed at the left periphery of a clause, with a co-referent pronoun in the core of the clause, e.g. "Ugandan radio stations they do mix their language and English" (dislocated element: "Ugandan radio stations"; co-referent pronoun: "they"). Interviews with the seven informants were transcribed and manually coded for features related to left-dislocation.

Biber et al (1999: 957) reported that the frequency of left-dislocation in Standard English in their data was "over 200 per million words". In the present dataset, the frequency ranged from 1,548 per million words to 7,598 per million words, and it is therefore clear that the research question can be answered in the affirmative. As regards the form of the dislocated element, the findings were more surprising, in that two structures were found which have not been extensively discussed in the literature, namely cases where the dislocated element is either a personal pronoun + a reflexive pronoun ("I myself I used to love this language"), or just a personal pronoun. The functions of left-dislocations were also explored, and these results will be included in the paper presentation, while future studies are required to explore whether the speakers transfer their preference for left-dislocation into their Norwegian production, and how this is interpreted by their new language community.

References

Biber, D., S. Johansson, G. Leech, S. Conrad, & E. Finegan. 1999. *Longman Grammar of Spoken and Written English.* Harlow: Pearson Education Limited.

Makalela, L. 2007. "Nativization of English among Bantu Language Speakers in South Africa." *Issues in Applied Linguistics* 15:2, 129-147.

Meierkord, C. 2004. “Syntactic variation in interactions across international Englishes.” *English World-Wide* 25:1, 109-132.

Mesthrie, R. 1997. “A sociolinguistic study of topicalisation phenomena in South African Black English.” In E. W. Schneider (ed.), *Englishes around the world: Studies in honour of Manfred Görlach. Volume 2: Caribbean, Africa, Asia, Australasia.* Amsterdam/Philadelphia: John Benjamins Publishing, 119-140.

Nassenstein, N. 2016. “A preliminary description of Ugandan English.” *World Englishes* 35:3, 396-420. DOI: https://doi.org/10.1111/weng.12205

Schmied, J. 2006. “East African Englishes.” In B. B. Kachru, Y. Kachru, & C. L. Nelson (eds.), *The Handbook of World Englishes*. Oxford: Blackwell Publishing Ltd, 188-202.

Schmied, J. 2008. “East African Englishes (Kenya, Uganda, Tanzania): morphology and syntax.” In R. Mesthrie (ed.), *Varieties of English. Africa, South and Southeast Asia.* Berlin: De Gruyter, 451-471.

Ssempuuma, J. 2017. *Morphological and Syntactic Feature Analysis of Ugandan English: Influence from Luganda, Runyankole-Rukiga, and Acholi-Lango.* PhD dissertation, Ruhr-Universität Bochum.

Read this Policy: A Corpus-Based Analysis of Terms of Use Contracts

Tim Samples, tsamples@uga.edu
Caroline Kraczon
Katherine Ireland (University of Georgia), katherine.ireland@uga.edu

"You should read this policy in full" is the introduction to Twitter's privacy policy. For most individuals and the wider public, routine transactional agreements like community guidelines for applications and platforms are regularly ignored. Despite this, a fundamental element of U.S. contract law is the duty to read, and consumers find themselves legally bound by many contracts they have never actually read (Becher 2008; Benoliel & Becher 2019). This proposed work-in-progress report expands on previous research in corpus linguistics and law by analyzing a corpus of terms-of-use (TOU), community guidelines, intellectual property policies, and privacy contracts in different mobile application categories. The corpus contains just under 1 million tokens across 10 different application categories. Application categories include social media and communication, business, education, gaming, shopping, dating, finance, and others.

Previous literature has focused on much smaller datasets (Becher 2008; Benoliel & Becher 2019), other types of legal provisions and documents (Kretzschmar et al. 2004; Alasmary 2019; Tuggener et al. 2020), and automatic text classification (Chalkidis et al. 2019; Hendrycks et al. 2021). This work includes discussion on the corpus-building process, specifically regarding balance and representation, in addition to preliminary results of analysis. Primary aims are comparison of linguistic structures and keywords across application categories and by policy types. Social media platforms are of specific interest due to their popularity and far-reaching implications. The utility of corpus-based methods is underscored through the analysis of frequencies of individual tokens and bigrams (Hunston 2010; Römer 2012; Biber et al. 2021), keywords (Kretzschmar et al. 2004; Baker et al. 2019), and KWIC lines through the use of CQP (Evert and Hardie 2011; Evert 2021) with R packages tidytext (Silge and Robinson 2021) and polmineR (Blätte & Leonhardt 2019). Log-likelihood (Dunning 1993; Gabrielatos and Baker 2008), raw frequencies, and ratio frequencies are reported. Log-likelihood is especially pertinent for highlighting typical contexts of use (Brezina 2018). Stubbs notes that frequencies of individual tokens and combinations of tokens encode and reveal larger discourses of meaning (1995). Preliminary results highlight the prevalence of references to users, with most frequent tokens including *you* and *your* and contractual indicators like *services, context, terms, use,* and *account* also being highly frequent. Frequency distributions of bigrams also evidence similar findings, including references to individual users and technical vocabulary like *the services, these terms,* and *use of.* Keywords differ significantly by category of application and type of contract.

This interdisciplinary research underscores the critical implications related to the unilateral nature of TOUs for law and policy, in addition to the efficacy of corpus-based methods for understanding the linguistic patterns within application contracts.

References

Becher, Shmuel I. 2008. Asymmetric Information in Consumer Contracts: The Challenge That is Yet to Be Met. *American Business Law Journal* 45(4), 723-774.

Benoliel, Uri and Shmuel I. Becher. 2019. The Duty to Read the Unreadable. *Boston College Law Review* 60(8), 2255-2296.

Biber, Douglas, Susan Conrad, and V. Cortes. 2004. If You Look at…Lexical Bundles in University Teaching and Textbooks, *Applied Linguistics.* 371-405.

Biber, Douglas, Stig Johansson, Geoffrey N. Leech, Susan Conrad, and Edward Finegan. 2021. *Grammar of Spoken and Written English. John Benjamins Publishing.*

Blätte, A. and C. Leonhardt. 2019. PolmineR() package, v 0.8.0.

Brezina, V. 2018. *Statistics in Corpus Linguistics: A Practical Guide.* Cambridge University Press.

Chalkidis, Ilias, Ion Androutsopoulos, and Achilleas Michos. 2017. Extracting Contract Elements. *ICAIL 17: Association for Computing Machinery.*

Dunning, Ted. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics 19 (1): 61–74. https://www.aclweb.org/anthology/J93-1003.*

Evert, S. and A. Hardie. 2011. Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 conference,* University of Birmingham, UK.

Evert, Stefan, and the CWB Development Team. 2021. The IMS Open Corpus Workbench (CWB) Corpus Encoding and Management Manual. http://cwb.sourceforge.net/

Feinerer, I. and K. Hornik. 2019. tm: Text Mining Package. R package version 0.7-7.

Gabrielatos, Costas and Paul Baker. 2008. Fleeing, Sneaking, Flooding: A Corpus Analysis of Discursive Constructions of Refugees and Asylum Seekers in the UK Press, 1996-2005. *Journal of English Linguistics* 36(1).

Hardie, A. 2012. 'CQP Web: combining power, flexibility, and usability in a corpus analysis tool', *International Journal of Corpus Linguistics*, pp. 380-409. John Benjamins Publishing.

Hendrycks, Dan, Collin Burns, Anya Chen, and Spencer Ball. 2021. CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review. *35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks.*

Hunston, Susan. 2010. How Can a Corpus Be Used to Explore Patterns? In *Routledge Handbook of Corpus Linguistics.* Ed. Anne O'Keeffe and Michael McCarthy. London: Routledge Publishing.

Kretzschmar, W., C. Darwin, C. Brown, D. Rubin, D. Biber. 2004. Looking for the Smoking Gun: Principled Sampling in Creating the Tobacco Industry Documents Corpus. *Journal of English Linguistics* (32)1.

R Core Team 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria.

Römer, Ute. 2010. Establishing the phraseological profile of a text type: The construction of meaning in academic book reviews. *English Text Construction.* John Benjamins Publishing.

Scott, Mike. 2010. What can corpus software do? In *Routledge Handbook of Corpus Linguistics.* Ed. Anne O'Keeffe and Michael McCarthy. London: Routledge Publishing.

Scott, Mike and Christopher Tribble. 2006. *Textual Patterns: Key words and corpus analysis in language education.* John Benjamins Publishing Company: Amsterdam.

Silge, Julia, and David Robinson. 2021. Textmining with R: A Tidy Approach.

Stubbs, M. 1995. Collocations and cultural connotations of common words. *Linguistics and Education.* 370-390.

Tuggener, Don, Pius von Däniken, Thomas Peetz, and Mark Cieliebak. 2020. LEDGAR: A Large-Scale Multilabel Corpus for Text Classification of Legal Provisions in Contracts. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020).*

Wickham et al. 2019. Welcome to the tidyverse. *Journal of Open Source Software* (43)4, https://doi.org/10.21105/joss.01686

Once upon a time, there was a fairy tale … And it lived happily ever after: A contrastive corpus-based study of German and English fairy tale openings and closings

Christina Sanchez-Stockhammer (TU Chemnitz), Asya Yurchenko (TU Chemnitz)
christina.sanchez@phil.tu-chemnitz.de; asya.yurchenko@phil.tu-chemnitz.de

It is a popular belief universally acknowledged that all fairy tales begin with the formula "*Es war einmal"* in German and its equivalent "*Once upon a time"* in English. However, there is actually considerable variation between the beginnings of the fairy tales collected by the brothers Jacob and Wilhelm Grimm. The same holds true for their endings, as virtually none of the Grimms' fairy tales conclude with the famous formula expected by a general readership, "*Und sie lebten glücklich und zufrieden bis an ihr Lebensende."/"And they lived happily ever after*." Since this issue has not been approached from a quantitative perspective in the literature so far, this shows the necessity for an empirical corpus stylistic treatment.

In the present paper, we introduce the multi-parallel TransGrimm corpus, which is currently being compiled at TU Chemnitz from the German fairy tales by the Brothers Grimm and their English translations with the aim to contribute novel insights to stylistic and stylometric research. The present study has set out to address the issue of fairy tale beginnings and endings from a quantitative perspective using corpus-linguistic methods. Based on material used for the compilation of the corpus, we extracted the first and last sentences from all the fairy tales in the most widely read German edition of the Grimms' "*Kinder- und Hausmärchen"* (1857) and Margret Hunt's (1884) complete English translation of the same edition.

We provide an overview of the frequency of opening variants like "*There was once"* or "*Once on a time"* and of closing lines like "*And they lived for a long time"* or [...] "*happily until their death*." We furthermore analysed both samples for the most frequent collocations of "*lived"* and also conducted an n-gram analysis of the first and last lines as a particularly suitable measure characterising the fairy tale genre (Sanchez-Stockhammer 2020).

A classification of the first-named entities in the openings reveals 'king' to be the most frequent profession (18.6 % of all professions and 8.8% of first-named entities in English overall). The examination of the last lines shows that the verb "*lived"* is exclusively followed by positive collocates (like "*as happy as a woodlark"* or *"in great magnificence"*). A diachronic analysis of the phrase "*once upon a time"* using Google Ngrams and the COHA corpus suggests that, despite its occurrence in only 7.7% of the fairy tale openings in Hunt, it has become the entrenched form, experiencing constant growth in usage since the 1880s, whereas its remarkably similar competitor "*once on a time"* (14.9%) has slid into obscurity.

References

Grimm, J., & Grimm, W. (1857). *Kinder- und Hausmärchen* (7th ed.). Göttingen: Dieterichsche Buchhandlung.

Grimm, J., & Grimm, W. (1884). *Grimm's Household Tales* (M. R. Hunt, Trans.). London: Bell and Sons.

Sanchez-Stockhammer, C. (2020). The potential of multi-word units as measures of fairy-tale style in Schneewittchen (Snow-White) and its English translations. In L.

Fesenmeier, & I. Novakova (Eds.), *Phraséologie et stylistique de la langue littéraire* (pp. 305-327). Bern: Peter Lang.

Trump's Populist Rhetoric: A Corpus-Based Analysis of 'The People', and 'The Elite'

Julia Schilling (University of Hamburg)
julia.schilling@uni-hamburg.de

Populism is on the rise and a prominent feature of the political landscape. According to Mudde (2004: 542), "populist discourse has become mainstream in the politics of western democracies. Indeed, one can even speak of a populist Zeitgeist". A prominent example is the 45th President of the United States, Donald J. Trump. While his language is deemed unusual for a political leader by many (Enli 2017: 56), Hunston (2017) claims it is "the true language of populism". However, exhaustive quantitative linguistic analyses of Trump's Twitter discourse from 2009 onwards that systematically compare the use of populist elements like people-centrism, and anti-elitism are still sparse. I address this research gap through a systematic study of Donald Trump's Twitter discourse from 2009 to the suspension of his account in January 2021, with the overall purpose of comparing Trump's discourse to the language of other U.S. politicians on Twitter, specifically Democratic and Republican senators, in order to investigate whether Trump employs a more populist rhetoric in his tweets and whether his rhetoric changes over time.

Since there is no agreement on a fixed definition of populism, guided by Laclau's (2005) theory of empty signifiers, I argue that populism is a discursively constructed political rhetoric that is characterized by people-centrism and anti-elitism. While people-centrism depicts 'the people' as a monolithic entity who are portrayed as morally superior to 'the elite', anti-elitism likewise displays 'the elite' as a homogenous unit that is in turn presented as morally inferior to 'the people'. Moreover, 'the people' are understood as the rightful sovereign whose will the populist claims to represent, and 'the elite' are simultaneously blamed for the mistreatment of 'the people' (Mudde 2004: 543). Linguistically, this will be analyzed not only through the frequency of references to nouns belonging to the semantic classes PEOPLE and ELITE but also through the use of the definite article *the* and the possessive determiner *our* as indicators for the creation of monolithic entities, and the portrayal of the people as morally superior and the elite as morally inferior respectively as a result of the use of adjectives modifying the semantic classes PEOPLE and ELITE. In addition, the portrayal of 'the people' as sovereign, as well as the blaming of 'the elite' is also analyzed through the syntactic position of the previously mentioned semantic classes and the verb processes used.

Results indicate that the language of Trump's tweets indeed exhibits populist elements of people-centrism and anti-elitism more than the tweets of U.S. senators, with a peak in the lead up to his election as the 45th President of the United States, and, therefore, hints at a strategic use of populist elements in his discourse on Twitter, dependent on the specific audience and his political goals.

References

Enli, Gunn (2017) "Twitter as Arena for the Authentic Outsider: Exploring the Social Media Campaigns of Trump and Clinton in the 2016 US Presidential Election". *European Journal of Communication* 32 (1): 50-61.

Hunston, Susan (2017) "Donald Trump and the Language of Populism". *University of Birmingham*. https://www.birmingham.ac.uk/research/perspective/donald-trump-language-of-populism.aspx (date of retrieval: 15 January 2022).

Laclau, Ernesto (2005) “Populism: What’s in a Name?”. In: Panizza, Francisco (ed.) *Populism and the Mirror of Democracy*, 32-49. London: Verso.
Mudde, Cas (2004) “The Populist Zeitgeist”. *Government and Opposition* 39(4): 542-563.

Pandem-onium: Identifying keywords and phrases in British COVID-19 Twitter and newspaper discourse

Julia Schilling (University of Hamburg) & Robert Fuchs
julia.schilling@uni-hamburg.de; robert.fuchs.dd@gmail.com

The COVID-19 pandemic has had profound influence on daily life, leading to intense public debate and a lexical innovation across many languages. Although linguists quickly started to document and analyze COVID-19 discourse (Baines et al. 2021; Saraff et al. 2021), there is as yet no systematic analysis of the lexical items and discourse patterns that characterize British COVID-19 discourse. We address this research gap through a systematic comparative analysis of public discourse during the COVID-19 pandemic. Through a big data approach, we identify not just distinct keywords and phrases linked to the pandemic but also track their development over time and across regions.

As news and social media posts can offer an insight into and simultaneously influence the public's perception of the COVID-19 pandemic, our analysis focuses on discourse in regional and national British newspapers as well as on the social media platform Twitter. The starting point of the analysis is a contrastive keyword analysis of the discourse of every month of 2019 with its equivalents in 2020 and 2021, comparing pandemic with pre-pandemic discourse, while filtering out seasonal effects (e.g. discussion of *snow* in January). Our data comprises material from January 2019 to December 2021, with more than 50 million geotagged tweets from the UK and 10% of all articles from 51 national and regional British newspapers.

Rather than collecting newspaper articles and tweets based on a pre-existing list of keywords, we use a data-driven approach to identify COVID-19 related n-grams ($1 \leq n \leq 4$) for each month of the pandemic based on log likelihood and log ratio. We then assign these keywords to semantic fields such as COVID-19 NAMES (e.g. *Covid-19*, *SARS-CoV-2*), PUBLIC HEALTH INSTRUCTIONS (e.g. *self-isolation*, *quarantine, PPE)*), VACCINATION and PEOPLE/INSTITUTIONS (e.g. *NHS, Boris Johnson, Matt Hancock*) and examine their development over time using statistical measures such as the median, standard deviation and skewness of the distribution of n-gram frequencies over time.

This analysis yielded over 300 1-grams, 350 2-grams, 200 3-grams, and 100 4-grams related to the COVID-19 pandemic. Results indicate that the lexis of British COVID-19 discourse significantly varies not only over time, but also within semantic fields of discourse and across regions. Preliminary results also indicate that the discourse on COVID-19 in newspapers tends to focus more on the societal impacts of the pandemic, mentioning, for example, frontline workers, airlines and travel warnings, while the discourse on Twitter is centered more around individual perspectives as well as rules and restrictions governing individual behavior, such as face mask requirements and curfews.

Beyond the analysis of temporal and regional variation in the frequencies of individual n-grams, the final step of our analysis attempts to identify clusters of COVID-19 related n-grams with similar frequency distributions. Using the k-means algorithm, we thus identify clusters of keywords whose relative frequencies co-vary across time and space, and we attempt to explain these associations with respect to several extra-linguistic variables (such as local infection, hospitalization and reproduction rates, local political orientations and overall duration of the pandemic).

References

Baines, Annalise; Ittefaq, Muhammad & Mauryne Abwao (2021) "#Scamdemic, #Plandemic, or #Scaredemic: What Parler Social Media Platform Tells Us About COVID-19 Vaccine". *Vaccines* 9 (421), pp. 1-16.

Saraff, Sweta; Singh, Tushar & Ramakrishna Biswal (2021) "Coronavirus Disease 2019: Exploring Media Portrayals of Public Sentiment on Funerals Using Linguistic Dimensions". *Frontiers in Psychology* 12:626638.

Fluency in Asian Englishes: A Multivariate Corpus-Based Analysis of Indian and Sri Lankan English

Karola Schmidt (JLU Giessen)
karola.schmidt-1@anglistik.uni-giessen.de

The study at hand explores fluency in spoken Indian and Sri Lankan English. In particular, it investigates productive utterance fluency (Götz 2013; Segalowitz 2010), i.e. those lexical (e.g. discourse markers) and temporal variables (e.g. silent pauses) speakers have at their disposal when overcoming planning phases. Previous research has focused on learner varieties of English (e.g. Dumont 2018) and has shown that these strategies, or fluencemes (Götz 2013), differ significantly in frequency across native and learner varieties (e.g. Götz 2013). Their use is, among other factors, governed by different sociolinguistics variables such as speaker age and gender (e.g. Stubbe & Holmes 1995; Tottie 2011). Aside from individual case studies (e.g. Revis & Bernaisch 2020 on pausing; Lange 2009 on discourse markers), no systematic fluency research into South Asian Englishes has been conducted yet, even though it stands to reason that second-language speakers have the same fluencemes at their disposal to facilitate speech production. Since both Indian and Sri Lankan English are generally taken to be nativised varieties in their own right (e.g. Schilk 2011; Bernaisch 2015), one would expect notable differences in fluenceme frequencies between the varieties since patterns of productive fluency can also be expected to emerge as markers of the evolutions of individual postcolonial Englishes. In order to test this, the study discusses three groups of core fluencemes – discourse markers (e.g. *like, you know*), unfilled pauses, and filled pauses (e.g. *uh, uhm*) – with regard to the following research questions:

1. Are there cross-varietal differences in the frequencies of core fluencemes between Indian and Sri Lankan English?
2. To what extent can the frequencies of core fluencemes be predicted based on sociolinguistic characteristics like speaker gender, age, and occupation?

38,579 unfilled pauses, 10,646 discourse markers, and 9,477 filled pauses were extracted from the spoken parts of the Indian and Sri Lankan components of the International Corpus of English. Their frequencies per speaker were analysed with regard to the VARIETY, speaker GENDER, AGE, and OCCUPATION, as well as the contextual factor of DISCOURSE MODE (i.e. face-to-face conversation or phone call). Linear regression models including two-way interaction terms were fitted for each of the three fluencemes under scrutiny. The analysis shows significant differences in fluenceme frequencies between the two varieties. In Indian English, frequencies of unfilled and filled pauses are higher than in Sri Lankan English. In contrast, Sri Lankan English speakers on average produce more discourse markers than Indian English speakers do. This suggests that the fluencemes in question are markers of regionalised patterns of productive fluency, which could be taken to support the notion of pragmatic nativisation. Additionally, GENDER was a relatively steady predictor across the models, showing that men tend to produce more fluencemes than women.

References

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

Dumont, Amandine (2018): *A Corpus Study of Non-native and Native Speaker (Dis)fluency Profiles*. PhD Dissertation. Louvain-la-Neuve: Université Catholique de Louvain.
Götz, Sandra (2013): *Fluency in Native and Nonnative English Speech.* Amsterdam: John Benjamins.
Lange, Claudia (2009): "Where's the party, yaar!: Discourse particles in Indian English. *World Englishes – Problems, Properties and Prospects*. Ed. Thomas Hoffmann, and Lucia Siebers. Amsterdam & Philadelphia: John Benjamins. 207-226.
Revis, Melanie, and Tobias Bernaisch (2020): "The pragmatic nativisation of pauses in Asian Englishes." *World Englishes* 39.1: 135-153.
Schilk, Marco (2011): *Structural Nativization in Indian English Lexicogrammar*. Amsterdam: John Benjamins.
Segalowitz, Norman (2010): *Cognitive Bases of Second Language Fluency*. New York: Routledge.
Stubbe, Maria, and Janet Holmes (1995): "*you know*, *eh* and other 'exasperating expressions': An analysis of social and stylistic variation in the use of pragmatic devices in a sample of New Zealand English." *Language & Communication* 15.1: 63-88.
Tottie, Gunnel (2011): "*Uh* and *Um* as sociolinguistic markers in British English." *International Journal of Corpus Linguistics* 16.2: 173-197.

Network visualisations of linguistic relationships in large datasets: A case study exploring the context of *normal* in British English

Hanna Schmück (Lancaster University)
h.schmueck@lancaster.ac.uk

In the wake of many 'new normals' in linguistics, new tools to explore linguistic data are steadily being proposed and implemented. This study aims to showcase one of these novel approaches: large scale linguistic network visualisations and how they can shed light on the interplay between language production and language perception. For the purpose of demonstrating strengths and limitations of this approach we carry out a case study on the basis of the following research questions:

> What is the collocational embedding of the word *normal* in spoken British English? What shape might the associative embedding of the word *normal* in native BrE speaker's mental lexicon take?

In order to explore this, a multidisciplinary approach is taken that spans corpus linguistics, psycholinguistics and graph theory. The exact method employed to investigate word embeddings/context here is based on a custom-built python scripts used to pre-process and weight corpus data as well as word association data. First sentence-span tuples from the spoken BNC 2014 (Love et al, 2017) are extracted while retaining their directionality – an often underreported yet crucial property (Michelbacher et al., 2011; Gries; 2013, McConnell & Blumenthal-Dramé, 2019) – and the corresponding MI^2 scores are calculated. Then the Small World of Words word association database (SWOW; De Deyne et al, 2019) is processed and filtered to only contain responses by British participants for comparability with the BNC. The obtained collocations ($MI^2 \geq 10$) and association pairs (Association weight ≥ 1) are fed to Cytoscape (Shannon et al., 2003) via py4cytoscape. The visual representations of the BNC and SWOW-UK networks surrounding the word *normal* (Figures 1 and 2) are created on the basis of an edge-weighted spring directed layout (Kamada & Kawai, 1989) which roughly maps MI^2 scores and association weights onto the displayed distances between words. Lastly, a range of graph theoretical properties for both networks are extracted and interpreted.

The results indicate that both the complete SWOW UK network and the complete Spoken BNC 2014 network exhibit small world properties (Watts & Strogatz, 1998). Qualitatively speaking, the collocations surrounding *normal* in the spoken BNC 2014 are generally sparser and the resulting subnetwork is denser than the one emerging from word associations. In the BNC, key topics surrounding the term are personal relationships and looks. These are organised around a strong network core composed of discourse markers and frequent verbs such as *like*, *really*, *well* etc. The word association dataset, however, looks markedly different and almost fractures into five distinct topic areas: antonyms of normality, society, health, abstract notions of normality, and discourse surrounding drugs. This suggests discrepancies between the mental association processes surrounding abstract terms such as *normal* and usage of such terms in everyday conversation.

Lastly, further applications of this methodology are briefly remarked upon. These include optimising language pedagogy practices (Xiao & McEnery, 2006; Webb & Kagimoto, 2009) via assessing how central certain terms are to a learner's language network and exploring network cliques which is useful for lexicography (Gablasova et al., 2017; Simpson-Vlach & Ellis, 2010) and researching language change (Chen et al., 2018).

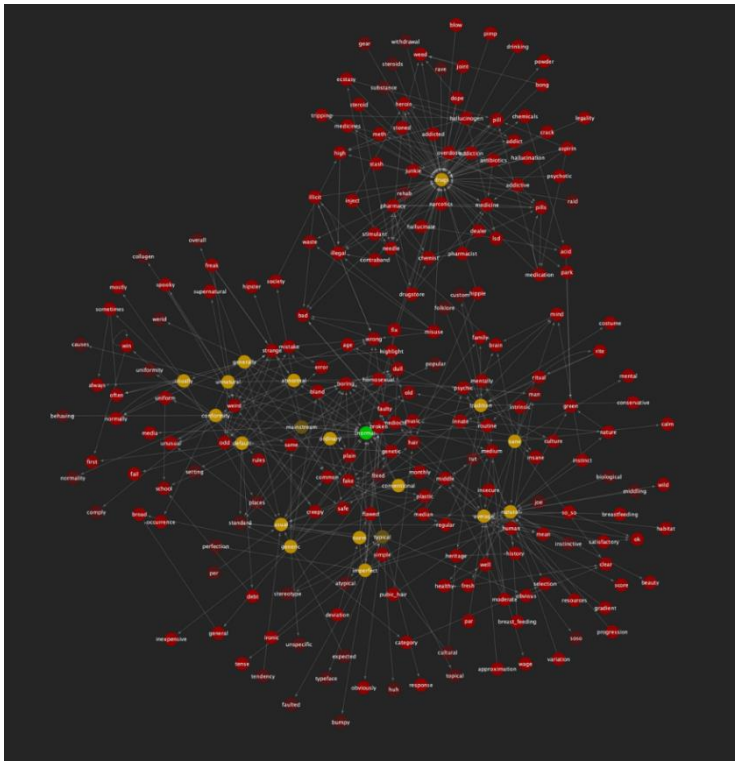

Figure 1: Edge-weighted spring directed network displaying the word association embedding of normal. Normal *marked in green, first order connections in yellow and second order connections in red.*

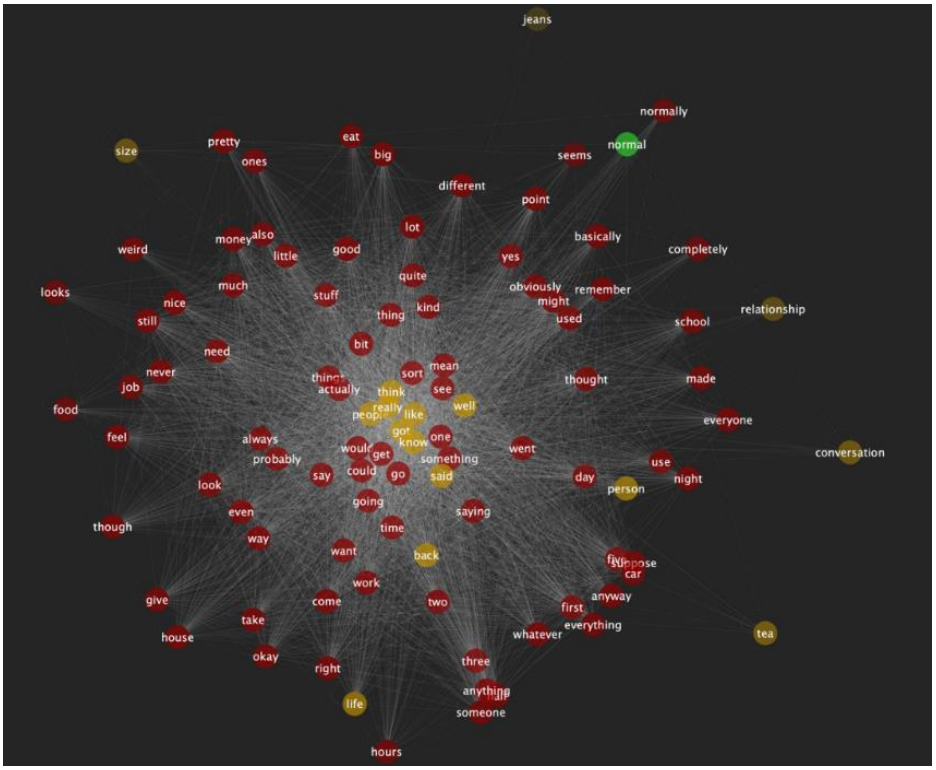

Figure 2: Edge-weighted spring directed network displaying the sentence collocation embedding of normal. Normal *marked in green, first order collocations in yellow and second order collocations in red.*

References

Chen, H., Chen, X., & Liu, H. (2018). How does language change as a lexical network? An investigation based on written Chinese word co-occurrence networks. PloS One, 13(2), e0192545. https://doi.org/10.1371/journal.pone.0192545

Deyne, S. de, Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2019). The "Small World of Words" English word association norms for over 12,000 cue words. *Behavior Research Methods*, *51*(3), 987–1006. https://doi.org/10.3758/s13428-018-1115-7

Gries, S. T. (2013). 50-something years of work on collocations: What is or should be next … *International Journal of Corpus Linguistics*, *18*(1), 137–166. https://doi.org/10.1075/ijcl.18.1.09gri

McConnell, K., & Blumenthal-Dramé, A. (2019). Effects of task and corpus-derived association scores on the online processing of collocations. *Corpus Linguistics and Linguistic Theory*, *aop*, 1–44. https://doi.org/10.1515/cllt-2018-0030

Michelbacher, L., Evert, S., & Schütze, H. (2011). Asymmetry in corpus-derived and human word associations. *Corpus Linguistics and Linguistic Theory*, *7*(2). https://doi.org/10.1515/cllt.2011.012

Gablasova, D., Brezina, V., & McEnery, T. (2017). Collocations in Corpus-Based Language Learning Research: Identifying, Comparing, and Interpreting the Evidence. Language Learning, 67(S1), 155–179. https://doi.org/10.1111/lang.12225

Kamada, T., & Kawai, S. (1989). An algorithm for drawing general undirected graphs. Information Processing Letters, 31(1), 7–15. https://doi.org/10.1016/0020-0190(89)90102-6

Love, R., Dembry, C., Hardie, A., Brezina, V. and McEnery, T. (2017). The Spoken BNC2014: designing and building a spoken corpus of everyday conversations. In *International Journal of Corpus Linguistics,* 22(3), pp. 319-344.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Research, 13(11), 2498–2504. https://doi.org/10.1101/gr.1239303

Simpson-Vlach, R., & Ellis, N. C. (2010). An Academic Formulas List: New Methods in Phraseology Research. Applied Linguistics, 31(4), 487–512. https://doi.org/10.1093/applin/amp058

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. Nature, 393(6684), 440–442. https://doi.org/10.1038/30918

Webb, S., & Kagimoto, E. (2009). The Effects of Vocabulary Learning on Collocation and Meaning. TESOL Quarterly, 43(1), 55–77.

Xiao, R., & McEnery, T. (2006). Collocation, Semantic Prosody, and Near Synonymy: A Cross-Linguistic Perspective. Applied Linguistics, 27(1), 103–129. https://doi.org/10.1093/applin/ami045

Separating the Wheat from the Chaff: how to detect Idioms vs. compositional Collocations with Collocation Measures and Distributional Semantics

Gerold Schneider (University of Zurich)
gschneid@ifi.uzh.ch

Collocation strength is traditionally calculated with frequentist approaches (Evert 2009, Pecina 2009, Gries 2013, Bartsch and Evert 2014) as semantic approaches (Wulff 2008) to the determination of collocations have been thought to perform less well. With the advance of distributional semantics (Baroni and Lenci 2010, Sahlgren 2006), it is time to reassess the situation. This study proposes to detect collocations by combining frequentist approaches and fixedness with measures from the word embedding space. Our research question is if and how much linearly combining collocation measures and distributional semantics improves the detection of collocations generally and non-compositional collocations (idioms) specifically.

As corpus material, we use the BNC (Aston & Burnard 1998) to calculate verb-PP collocations (Lehmann & Schneider 2011). We combine and filter the collocation values by distributional semantic measures trained on large web corpora (BNC, UKWac, and a Wikipedia dump, see Günther et al. 2014) with word2vec (Mikolov et al. 2013). We evaluate our results by means of a manual classification of the collocation lists by Lehmann & Schneider (2011), and by external resources (e.g. Bartsch & Evert 2014). The idea to use distributional semantics to distinguish between compositional and non-compositional word-combinations is used by Maldonado-Guerra & Emms (2011) and Salehi et al. (2015), who predict compositionality of frequent word sequences, or multi-word expressions from a given list. Unlike our approach, they do not consider collocation statistics, though, while we combine collocation measures with distributional measures.

For a collocation candidate composed of the words *w1* and *w2* we calculate:

- the cosine distance in the semantic space between *w1* and *w2*
- shared synonyms between synonyms of *w1* and synonyms of *w2*
- how often words semantically close to *w1* appear in its stead

For example, the cosine similarity between *pale* and *significance*, which features in the non-compositional collocation *pale into insignificance* is 0.24, while the similarity between *plug* and *socket*, as in the compositional collocation *plug into socket*, is 0.87. Also when we compare collocations with the same *w1*, we find, for instance, that *fall* and *disrepair* (*fall into disrepair*) are less similar than *fall* and *trap* (*fall* into *trap*).

Related approaches combining collocation detection with distributional semantics can be found in several experiments. For example, Wanner et al. (2016) classify collocations into different semantic fields with distributional semantics. Ljubešić et al. (2021) report a small increase in collocation detection for Slovenian based on distributional semantics, but they do not consider compositionality. The parsimonious linear combination of the two approaches, which we propose here, is missing so far.

Preliminary results show considerable improvement over frequentist approaches, particularly when it comes to idiomatic expressions, which are non-compositional collocations, and for instance particularly useful for teaching English, or for compiling dictionaries of fixed expressions.

In a nutshell, we put to the test if adding distributional semantics manages to separate the wheat from the chaff, or if it is a case of throwing good money after bad. Our

detected idioms and *R* code will be made publicly available online to ensure reproducibility.

References

Aston, Guy and Burnard, Lou. 1998. The BNC Handbook. Exploring the British National Corpus with SARA. Edinburgh University Press, Edinburgh .

Baroni, Marco and Lenci, Alessandro. 2010. "Distributional Memory: A general framework for corpus-based semantics". *Computational Linguistics*, 36, 4, 673-721.

Bartsch, Sabine and Evert, Stefan. 2014. "Towards a Firthian notion of collocation". *Vernetzungsstrategien, Zugriffsstrukturen und automatisch ermittelte Angaben in Internetwörterbüchern*, 2/2014, 48--61.

Evert, Stefan. 2009. "Corpora and collocations". *Corpus Linguistics. An International Handbook*, article 58, 1212-1248.

Gries, Stefan. (2013). 50-something Years of Work on Collocations. *International Journal of Corpus Linguistics*, *18*(1), 137–165.

Günther, Fritz, Dudschig, Carolin, and Kaup, Barbara. 2014. "LSAfun - An R package for computations based on Latent Semantic Analysis". *Behavior research methods*, 47.

Lehmann, Hans Martin and Schneider, Gerold. 2011. "A large-scale investigation of verb-attached prepositional phrases". Studies in Variation, Contacts and Change in English, Volume 6: *Methodological and Historical Dimensions of Corpus Linguistics*. Helsinki: VariEng.

Ljubešić, Nikola, Logar, Nataša, and Kosem, Iztok. 2021. "Collocation ranking: frequency vs semantics." *Slovenščina 2.0: empirical, applied and interdisciplinary research* 9.2, 41-70.

Maldonado-Guerra, Alfredo and Emms, Martin. 2011. "Measuring the compositionality of collocations via word co-occurrence vectors: shared task system description". In *Proceedings of the Workshop on Distributional Semantics and Compositionality (DiSCo '11)*. Association for Computational Linguistics, USA, 48–53.

Mikolov,Tomas, Chen, Kai, Corrado, Greg, and Dean, Jeffrey. "Efficient estimation of word representations in vector space". *CoRR*, abs/1301.3781, 2013.

Pecina, Pavel. 2009. *Lexical Association Measures: Collocation Extraction*. Institute of Formal and Applied Linguistics, Charles University in Prague, Czech Republic .

Sahlgren, Magnus. 2006. *Word-Space Model: Using distributional Analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. PhD thesis, University of Stockholm.

Salehi, Bahar, Cook, Paul, and Baldwin, Timothy. 2015. "A Word Embedding Approach to Predicting the Compositionality of Multiword Expressions". In *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, 977–983.

Wanner, Leo, Ferraro, Gabriela, and Moreno, Pol. (2017). Towards Distributional Semantics-Based Classification of Collocations for Collocation Dictionaries. *International Journal of Lexicography*, *30*(2), 167–186.

Wulff, Stefanie. 2008. *Rethinking Idiomaticity*. London: Continuum.

A negator walks into a modal clause
A diachronic corpus-based study of a complex relationship

Ulrike Schneider (University of Mainz)
ulrike.schneider@uni-mainz.de

The relationship between modals and negation has attracted a lot of attention in past decades. Studies mostly focus on logic, i.e. on the scope of the negator or the modal (cf. e.g. De Haan 1997; Palmer 1997; Radden 2009) and on how scope can be modelled within generative frameworks (cf. also Roberts 1985; Beukema and van der Wurff 2002; Hankamer 2011). To date, there are few empirical – let alone diachronic – studies on the relationship between modals and negation (cf. e.g. Bergs 2008; Daugs 2021). Consequently, there are still many open questions concerning the frequency of negated modal verb phrases, their dominant contexts of use and how these may have changed over time.

What we do know is that some modals strongly attract negation while others do not. In Present-Day English *can* and *could*, for instance, are far more likely to co-occur with negation than other core modals (cf. Mindt 1995, see also Römer 2004). Furthermore, across a variety of studies on verbal constructions (Schneider 2021, 2022), the same correlation appeared time and again: If the verb under investigation was accompanied by a modal, the chance that it was also negated was significantly higher than in non-modal uses of the same construction. The strength of the correlation varied, though; in some cases, rates were only marginally higher while in others they increased by a factor of ten (e.g. *force*). Yet the correlation surfaced in American and British data, in Early Modern and Modern English and irrespective of whether the construction attracted modals (e.g. causative *bring*) or whether it did not (e.g. causative *make*). In many contexts, the attraction increased in the 20th century; in the case of causative *bring* even up to a point where modals almost function as negative polarity items in the construction.

The present study sheds light on modal-negated clauses in English by addressing the following questions:

1. Did the replacement of 'V *not*' by 'Aux *not* V' in Early Modern English as well as the extensive changes in the field of modals occurring in the 19th and 20th centuries (cf. e.g. Leech 2013) lead to changes in co-occurrence patterns?
2. Does the increasing subjectification in the writing style of modern authors (cf. e.g. Verhagen 2000) have an influence on the combined use of modals and negation?
3. Are there further factors, like lexical aspect, which promote the combined use of modals and negation?

For the purpose, I extracted all tokens of core modals from the Chadwyck Healey collection as well as from the wridom1 subcorpus of the BNC. The resulting 1.4-million-word dataset covers modal use in British prose published between ca. 1500 and 1990. It is complemented by a second dataset comprised of all 700,000 instances of *not/n't* occurring in the same corpora. These datasets will be analysed with the help of Configural Frequency Analysis as well as distinctive collexeme analysis (cf. Gries and Stefanowitsch 2004; Hilpert 2016).

References

Bergs, Alexander. (2008). *Shall* and *Shan't* in Contemporary English – a Case of Functional Condensation. In Trousdale, Graeme and Nikolas Gisborne (Eds.): *Constructional Approaches to English Grammar*. Berlin: De Gruyter Mouton. 113-143.

Beukema, Frits and Wim van der Wurff. (2002). Modals, Objects and Negation in Late Middle English. In Barbiers, Sjef, Frits Beukema and Wim van der Wurff (Eds.): *Modality and Its Interaction with the Verbal System*. Amsterdam/Philadelphia: John Benjamins. 75-102.

Daugs, Robert. (2021). Contractions, Constructions and Constructional Change. Investigating the Constructionhood of English Modal Contractions from a Diachronic Perspective. In Hilpert, Martin, Bert Cappelle and Ilse Depraetere (Eds.): Modality and Diachronic Construction Grammar. Amsterdam/New York: John Benjamins. 13-51.

De Haan, Ferdinand. (1997). *The Interaction of Modality and Negation*. London/New York: Routledge.

Gries, Stefan Th. and Anatol Stefanowitsch. (2004). Extending Collostructional Analysis. A Corpus-Based Perspective on 'Alternations'. *International Journal of Corpus Linguistics* 9(1): 97-129.

Hankamer, Jorge. (2011). Auxiliaries and Negation in English. In Gutiérrez-Bravo, Rodrigo, Line Mikkelsen and Eric Potsdam (Eds.): *Representing Language. Essays in Honor of Judith Aissen.* Santa Cruz: University of California, Linguistics Research Center. 121-135.

Hilpert, Martin. (2016). Change in Modal Meanings. Another Look at the Shifting Collocates of May. *Constructions and Frames* 8(1): 66-85.

Leech, Geoffrey. (2013). Where Have All the Modals Gone? An Essay on the Declining Frequency of Core Modal Auxiliaries in Recent Standard English. In Marín-Arrese Juana, I., Marta Carretero, Jorge Arús Hita and Johan van der Auwera (Eds.): *English Modality. Core, Periphery and Evidentiality.* Berlin/Boston: Mouton de Gruyter. 95-115.

Mindt, Dieter. (1995). *An Empirical Grammar of the English Verb: Modal Verbs.* Berlin: Cornelsen.

Palmer, Frank. (1997). Negation and Modality in Germanic Languages. In Swan, Toril and Olaf J. Westvik (Eds.): *Modality in Germanic Languages: Historical and Comparative Perspectives*. Berlin: Mouton de Gruyter. 133-149.

Radden, Günter. (2009). Affirmative and Negated Modality. *Quaderns de Filologia. Estudis Lingüístics.* 14: 169-192.

Roberts, Ian. (1985). Agreement Parameters and the Development of English Modal Auxiliaries. *Natural Language & Linguistic Theory* 3: 21-58.

Römer, Ute. (2004). A Corpus-Driven Approach to Modal Auxiliaries and Their Didactics. In Sinclair, John (Ed.): *How to Use Corpora in Language Teaching*. Amsterdam/Philadelphia: John Benjamins. 185-202.

Schneider, Ulrike. (2021a). Loss of Intersective Gradience as the Lifeboat of a Dying Construction. An Analysis of the Diachronic Change of Causative *bring*. *Folia Linguistica Historica* 42(2), 429-59.

Schneider, Ulrike. (2022). *They're proing it up hardcore*. An Analysis of the V *it up* Construction. In Matthias Eitelmann & Dagmar Haumann (eds.): *Extravagant Morphology. Studies in Rule-Bending, Pattern-Extending and Theory-Challenging Morphology.* 207-32. Amsterdam/Philadelphia: John Benjamins.

Verhagen, Arie. (2000). Interpreting Usage: Construing the History of Dutch Causal Verbs. In Barlow, Michael and Suzanne Kemmer (Eds.): *Usage-Based Models of Language*. Stanford: CSLI.

Compiling a diachronic corpus to trace variety-specific genre conventions acrosstime: challenges and solutions for automatising text recognition of businesscorrespondence from Hong Kong

Ninja Schulz (University of Würzburg), ninja.schulz@uni-wuerzburg.de
Lisa Lehnen (University of Würzburg), lisa.lehnen@uni-wuerzburg.de
Christian Reul & Carolin Biewer

Compiling long called-for diachronic corpora for outer circle varieties of English, though well underway by now, still poses a problem: The text types preserved are not necessarily those that would seem most interesting for world Englishes research. In the case of Hong Kong, being a financial and commercial hotspot, business correspondence has been in place from early onwards. In the mid-19th century, banks and companies, such as the Hong Kong Shanghai Banking Corporation (HSBC) and Jardine Matheson & Co. (JM), established their head offices in Hong Kong. Business correspondence, however, constitutes a highly formal and conventionalised genre (though internal stylistic variation reflects a spectrum from private and confidential letters to official letters sent to external addressees), which was initially dominated by writers whose native language was British English of the late 19th century. Although local Hong Kong staff has been gradually employed and trained for business (Hao, 1982, pp. 86–87; HSBC, n.d.; Life at HSCB, 2016) since the beginnings of the 20th century, nativised structures in the form of variety-specific morphosyntactic features cannot necessarily be expected. What can be expected though are changes regarding expressions of stance, power and deference, as well as innovations in the distribution of constructions, preferred formulaic language, and lexical bundles as "PCEs are characterized by the emergence of new constructions, new habits of word combinations which are meaningful (only) in a given speech community (Schneider, 2009, p.88). For a systematic analysis of the second group of features, large amounts of data are needed. Although many business letters from Hong Kong were preserved, the quality of the material poses a challenge to its digitisation and preparation for corpus linguistic analyses.

This work in progress report explores the methodological issues of processing these historical documents by automatising as much of the workflow as possible. So far, we have collected about 4,000 business letters written in Hong Kong from as early as the 1860s until the beginnings of the 1970s. The focus lies on the automated text recognition (not the annotation) of the business letters, which, though mostly typed since the 1920s, are often heterogeneous in terms of layout and quality (regarding the scan as well as the condition of the original document). In collaboration with the Centre for Philology and Digitality (ZPD) at the University of Würzburg, we use the open-source software OCR4all, which was especially designed to handle historical printings and manuscripts. We implemented the workflow whose final stage is the manual correction of the recognised text on a small portion of the material (i.e. 125 letters) to train a domain-specific OCR-model with the aim of eventually managing larger amounts of data while keeping manual corrections as well as error rates at a minimum. While reporting the current challenges encountered, we will also give an outlook on how even handwritten letters can be automatically processed by the software in the future.

References

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

All-cleft constructions in the London–Lund Corpus 2 (LLC–2) of spoken British English

Eleni Seitanidi (Lund University), eleni.seitanidi@englund.lu.se
Nele Põldvere & Carita Paradis

This study contributes to the hitherto limited research on *all*-cleft constructions (*all you have to do is flip it*). *All*-cleft constructions comprise two clauses connected with the copula *be*, where one clause is introduced by *all*, which is synonymous with 'only' (Traugott, 2008). The construction often conveys an assertion of the scarcity or inadequacy of what is discussed (Homer, 2019; Tellings, 2020). Given the construction's evaluative character (Traugott, 2008: 9), we argue that speakers use the 'not much' meaning of the construction to construe the action modified by the modal (when there is one) and the main verb of the *all*-clause as either (i) positive, e.g. a request by the speaker as causing minimal inconvenience (*all I need to know is do you want*), or (ii) negative, e.g. the action taken by the addressee or someone else as being inadequate (*all he has to do is forward an email to the solicitor and just hasn't*). We use data from the new London-Lund Corpus (LLC–2) (Põldvere et al., 2021) of spoken British English, and we adapt Biber at al.'s (1999) classification of *wh*-cleft constructions and apply it to *all*-clefts, categorizing *all*-cleft variants into (1) regular (*all I did is I wrote to them*), (2) reversed (*hydrogen is all I have available*), and (3) demonstrative (*be an adult that's all I'm asking you to do*) *all*-cleft constructions.

Adopting a Construction Grammar approach, we account for the (1) formal features, i.e., collocational patterns involving modal and main verbs, and intonation placement and contours, (2) meanings and (3) discourse functions of the *all*-cleft construction addressing the following research questions:

1. What are the form-meaning characteristics of the construction?
2. What are the communicative functions of the construction?
3. How do the communicative functions of the construction vary by discourse context, e.g. spontaneous commentary?

We combine quantitative and qualitative methods using concordances and word lists to study the formal features of the construction. We conduct close analyses of the texts to study the construction's communicative functions in each discourse context. Furthermore, the availability of the corpus audio files allows us to study the intonational patterns of the *all*-cleft constructions using specialised software.

Some preliminary findings suggest that proper *all*-clefts are the most frequent, followed by demonstrative *all*-clefts, and reversed *all*-clefts. With regard to the main verbs found in the construction, the majority express ACTION, i.e. *do*, followed by MENTAL STATES, i.e. *need*, *want*, *take*, then SPEAKING, i.e. *say*, and MENTAL ACTIVITY, e.g. *think*, *know*. In terms of the distribution of the construction across the discourse contexts of the corpus, preliminary findings suggest that the construction most frequently occurs in spontaneous commentary, followed by prepared speech and distanced conversations, face-to-face conversation, and, finally, legal proceedings. The distribution may be due to the fact that spontaneous commentary comprises monologues from cooking and science demonstrations where speakers often use the construction to explain the procedures and present them as straightforward.

References

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. (1999) *Longman Grammar of Spoken and Written English*, Harlow: Pearson Education.

Homer, V. (2019) That's all, in R. Stockwell, M. O'Leary, Z. Xu, and Z.L. Zhou, (eds.) *Proceedings of the 36th West Coast Conference on Formal Linguistics*, 1-21, Somerville: Cascadilla Proceedings Project.

Kay, P. (2013) The Limits of Construction Grammar, in G. Trousdale and T. Hoffmann (eds.) *The Oxford Handbook of Construction Grammar*, New York: Oxford University Press.

Põldvere, N., Johansson, V., and Paradis, C. (2021) On the London–Lund Corpus 2: Design, challenges and innovations, *English Language and Linguistics, 25*(3), 459-483.https://doi.org/10.1017/S1360674321000186

Tellings, J. (2020) An analysis of all-clefts, *Glossa: A Journal of General Linguistics*, *5*(1), 125.

Traugott, E., C. (2008) "All that he endeavored to prove was…": On the emergence ofgrammatical constructions in dialogual and dialogic contexts, in R. Cooper and R. Kempson (eds.) *Language in Flux: Dialogue Coordination, Language Variation, Change and Evolution*, pp: 143-177, London: Kings College Publications.

Stating the obvious: assumed evidential parentheticals with *verba dicendi* in contemporary Englishes

Mario Serrano-Losada (Complutense University of Madrid), mario.serrano@ucm.es
Zeltia Blanco-Suárez (University of Santiago de Compostela), zeltia.blanco@usc.es

The present paper focuses on two (near-)synonymous constructions of assumed evidentiality in English, the expressions *needless to say* and GO *without saying* (cf. examples (1) and (2)), which have so far remained relatively understudied (cf. Blanco-Suárez & Serrano-Losada 2017; Schmid 2020).

(1) **Needless to say**, this immediately provoked a scandal. (BYU-BNC, 1992)
(2) I still like him. I mean, I love him – **that goes without saying**. (COCA, 2006)

According to general dictionaries of English both expressions are (near-)synonyms of evidential adverbs like *obviously*, *of course* and *evidently*. Over the course of history, the constructions under study have become fixed expressions and are often used parenthetically to express non-propositional, procedural meaning. Alongside their use as hedging devices to shade categorical assertions, they also serve an evidential function, indicating assumption, logical reasoning or general knowledge (Aikhenvald 2004: 63).

The paper has a twofold aim. On the one hand, it describes and contrasts both constructions in Present-day American and British English with the purpose of establishing the functions and usage of these competing forms across registers. On the other hand, it analyzes their variation across several varieties of World Englishes by looking at the use and distribution of *needless to say* and GO *without saying* in five different varieties, two belonging to the 'inner circle', namely South African English and Canadian English, and three to the 'outer circle', Indian, Singapore and Philippine English (cf. Kachru 1985). These five varieties also differ as regards their norm-providing matrilects (either British or American English). The results of this contrastive study will thus shed light on the frequency and functions of *needless to say* and GO *without saying* in these varieties as compared to their prevalence and usage in the norm-providing ones.

Data for this paper have been drawn from the *Brigham Young University-British National Corpus* (BYU-BNC), the *Corpus of Contemporary American English* (COCA) and the *Corpus of Global Web-Based English* (GloWbE). The resulting datasets were annotated according to morphosyntactic, semantic, and structural criteria (e.g. parenthetical vs clausal).

The preliminary results indicate that *needless to say* is the expression with the widest variety of functions in contemporary British and American English, being mostly used as a parenthetical in the left periphery with forward scope. GO *without saying*, in turn, is mainly attested in extraposed clauses at the right periphery. These findings are partly replicated in the data from the norm-developing varieties: while the overall frequencies in Canadian, Philippine, and Indian Englishes pattern closely after those of the norm-providing varieties (American and British English), Singapore and South African Englishes clearly outweigh the reference varieties. Ultimately, this study aims to provide further insight into the patterns of usage of evidential discourse-pragmatic expressions across varieties of English (cf. Leuckert & Rüdiger 2021).

Sources

BYU-BNC = *Brigham Young University-British National Corpus*, compiled by Mark Davies. 2004-. Available online at: http://corpus.byu.edu/bnc/.

COCA = *The Corpus of Contemporary American English*, compiled by Mark Davies. 2008-. Available online at: http://corpus.byu.edu/coca/.

GloWbE = *The Corpus of Global Web-based English: 1.9 billion words from speakers in 20 countries*, compiled by Mark Davies. 2013. Available online at: https://corpus.byu.edu/glowbe/

References

Aikhenvald, Alexandra Y. 2004. *Evidentiality*. Oxford: Oxford University Press.

Blanco-Suárez, Zeltia & Mario Serrano-Losada. 2017. The rise and development of parenthetical *needless to say*: An assumed evidential strategy. *Journal of Historical Linguistics* 7(1/2). 134–159.

Kachru, Braj B. 1985. Standards, codification and sociolinguistic realism: The English language in the outer circle. In Randolph Quirk & Henry George Widdowson (eds.), *English in the world: Teaching and learning the language and literatures*, 11–30. Cambridge: Cambridge University Press for The British Council.

Leuckert, Sven & Sofia Rüdiger. 2021. Discourse markers and world Englishes. *World Englishes* 40(4). 482–487.

Schmid, Hans-Jörg. 2020. *The Dynamics of the Linguistic System: Usage, Conventionalization, and Entrenchment*. Oxford, New York: Oxford University Press.

Compiling a corpus of South Asian online Englishes: Some reflections and a pilot study

Muhammad Shakir (WWU Muenster) & Dagmar Deuber
muhammadshakiraziz@outlook.com

In this paper, we aim to present a new dataset of internet-based registers that is currently in-compilation and a pilot study based on this dataset. The use of the internet as a source for data collection in English Linguistics and World Englishes is obviously not new, the most common example being GloWbE (Davies and Fuchs, 2015) and other datasets available from, for example, www.English-corpora.org. However, we believe that smaller carefully compiled datasets have their own place due to potential problems with bigger datasets, for example see Loureiro-Porto (2017) for a comparison of GloWbE and ICE corpora. At the same time, we do not follow the ICE add-on corpora approach presented by Kirk and Nelson (2018) due to scarce availability of certain spoken internet-based genres, e.g. video blogs from S. Asian countries like Bangladesh, and the time and labour required for transcription of such genres.

The corpus consists of five subcategories, i.e. text messages, blogs and general websites, reader comments from news websites, discussion forums and tweets, originating from 6 countries, i.e. Bangladesh, India, Pakistan, Sri Lanka, the UK and the USA. Each subcategory consists of 1 million words. Text messages are an exception because only about 500,000 words have been collected from the four South Asian (SA) countries due to hurdles like the accessibility to such data and the time and effort required for annotation. The data for the other four genres were programmatically downloaded from their respective sources after verifying their country of origin. Text files were annotated and/or edited for indigenous/foreign content, copied content, potentially spam content, computer code etc.

In the pilot study we show how the Hindi/Urdu discourse marker *yaar* and tag question *na* (Lange, 2009) are not only used in India and Pakistan but in Bangladesh and Sri Lanka too. In India and Pakistan these discourse level items occur in all interactive genres but more frequently in text messages and tweets. In Bangladesh and Sri Lanka they generally occur in tweets highlighting the online language contact among these countries. For example, in our data "drop a selfie *na*" was used by a Bangladeshi fan of an Indian singer on Twitter, which shows the cultural and linguistic influence of Indian entertainment industry on other SA countries. Additionally, we discuss the Sri Lankan English tag question *neh* that was pointed out by one of our informants. *Neh* is almost exclusive to the Sri Lankan part of our corpus and most frequently occurs in text messages. The pilot study, thus, underscores the importance of corpus annotation for multilingual resources and the use of indigenous informants in this process.

Lastly, we discuss the challenges presented by each genre in terms of source verification and tagging, especially tagging for copied content, issues like tagging inconsistencies and possible areas of improvement.

References

Davies, M., & Fuchs, R. (2015). Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English Corpus (GloWbE). *English World-Wide*, *36*(1), 1–28.

Kirk, J., & Nelson, G. (2018). The International Corpus of English project: A progress report. *World Englishes*, *37*(4), 697–716. https://doi.org/10.1111/weng.12350

Lange, C. (2009). 'Where's the party *yaar*!' Discourse particles in Indian English. In T. Hoffmann & L. Siebers (Eds.), *World Englishes–problems, properties and prospects. Selected papers from the 13th IAWE conference* (pp. 207–222). Benjamins.

Loureiro-Porto, L. (2017). ICE vs GloWbE: Big data and corpus compilation. *World Englishes*, *36*(3), 448–470. https://doi.org/10.1111/weng.12281

Attitudes to immigration in Australian Hansard: 1970-2020

Adam Smith (Macquarie University) & Minna Korhonen (Macquarie University)
adam.smith@mq.edu.au

Previous research into the representation of immigration and immigrants in the public forum has used both qualitative discourse analysis and quantitative corpus analysis to investigate attitudes displayed in data from newspapers and parliamentary interactions. While Greenslade (2005) provides an historical perspective on the coverage of immigration issues in the British press – with case studies from the 1940s to the 2000s – much of the research done has been looking at recent shifts of attitude due to contemporary refugee issues across Europe. Gabrielatos & Baker (2008) demonstrate different levels of negativity presented by common collocations with refugees and asylum seekers in UK newspaper data, 1996-2005, while Walter (2002) contrasts positive with negative representations of cultural diversity in mass media across Europe, 1995-2000. More recently, Fotopoulos & Kaimaklioti (2016) provide quantitative corpus evidence to suggest that the media coverage of the Syrian crisis across Greece, Germany and Britain is generally positive towards the refugees. There are also examples of parliamentary discourse used as data, notably an edited volume by Wodak and van Dijk (2000) looking at examples on ethnic issues from parliamentary records in six European states.

In Australia, there has been much discussion of attitudes towards immigration in terms of social justice (e.g. Ghezelbash, 2018; Jupp 2009) and some analyses of the discourse around the topic (e.g. Martin 2021; Morrisey & Schalley 2017), but no systematic, diachronic review of the language used to represent refugees, immigrants and asylum seekers. Australia is a particularly interesting focus for discussion of changing attitudes as it has historically relied so heavily – economically and socially – on immigration (Castles, 2009). In order to chart the changes of attitude over a range of historical and political movements, this study has undertaken to collect data from the last 50 years, representing a period from before the acceptance of the first boat people from Vietnam to the current national policy of prevention symbolised by strict policing of the sea borders, and offshore detention of asylum seekers. We have collected a corpus of 28 million words from Parliamentary Hansard, composed of debates and speeches from the period 1970-2020 where immigration is the main, or a significant topic. These texts were identified by searching in the records using a set of relevant keyterms identified from the general literature, including *asylum seeker, emigrant, foreigner, immigrant*, as well as terms that have had a special impact in Australia like *boat people* and *queue jumper*. Irrelevant hits were removed and metadata was downloaded with the texts to provide biodata about the speakers involved.

Preliminary findings demonstrate variability across time in the portrayal of immigrants to Australia, both in the use of emotive terms such as *boat people* and *queue-jumper*, and positive/negative collocations with more official terms like *asylum seeker* and *refugee*. Contrastive attitudes are also evident across the two parliamentary houses (the House of Representatives and the Senate) and in the way that members of the governing party and the opposition parties address the issue.

References

Castles, S. (1992). Australian multiculturalism: social policy and identity in a changing society. In G. Freeman & J. Jupp 1992 *Nations of Immigrants*. Oxford University Press

Fotopoulos, S. & Kaimaklioti, M. (2016). Media discourse on the refugee crisis: on what have the Greek, German and British press focused? *European View* 15:265–279.
Ghezelbash, D. (2018). *Refuge Lost: Asylum Law in an Interdependent World*. Cambridge University Press.
Gabrielatos, C. & Baker, P. (2008). Fleeing, Sneaking, Flooding: A Corpus Analysis of Discursive Constructions of Refugees and Asylum Seekers in the UK Press, 1996-2005. *Journal of English Linguistics*, 36:1, 5-38.
Greenslade, R. (2005). Seeking scapegoats: The coverage of asylum in the UK press. *Asylum and Migration Working Paper 5*. Institute for Public Policy Research.
Jupp, J. (2009). *From White Australia to Woomera*. Cambridge University Press.
Martin, C. (2021). Jumping the queue? The queue-jumping metaphor in Australian press discourse on asylum seekers. *Journal of Sociology* 57(2): 343-361.
Morrisey, L. & Schalley, A.C. (2017). A Lexical Semantics for Refugee, Asylum Seeker and Boat People in Australian English. *Australian Journal of Linguistics* 37(4): 389-423.
Walter, J. (2002). Racism and cultural diversity in the mass media: An overview of research and examples of good practice in the EU Member States, 1995-2000. European Research Centre on Migration and Ethnic Relations (ERCOMER).
Wodak, R. & van Dijk, T.A. (eds.) (2000). *Racism at the Top: Parliamentary discourses on ethnic issues in six European states*. Drava Verlag.

Keyword analysis: Progress through regression

Lukas Sönning (University of Bamberg)
lukas.soenning@uni-bamberg.de

The purpose of a keyword analysis is to identify in a corpus of interest those (lexical) elements that are overrepresented relative to a reference corpus. How to go about identifying keywords has been subject to methodological debate. The default statistical maneuver, a likelihood ratio or chi-square test, has met with justified criticism due to its statistical inadequacy (Kilgarriff 2005; Bestgen 2014; Koplenig 2019). Some have rightly argued that keyness should be expressed in descriptive terms, to reflect the actual degree to which an item is overrepresented (Hardie 2014; Brezina 2014). Finally, text-level analyses have gone further still by accounting for the fact that corpora consist of text samples (e.g. Brezina & Meyerhoff 2014; Lijffijt et al. 2016).

This paper introduces a method that unifies methodological advances in keyword analysis. It taps into the family of count regression models (Cameron & Trivedi 2013); specifically, negative binomial regression. This form of Poisson regression accommodates and quantifies overdispersion, which corresponds to the corpus linguistic notions of dispersion and "burstiness". Previous methodological developments resonate in this technique, which (i) adopts a text-level analysis and (ii) measures keyness in a transparent and descriptive way, namely as a ratio: the rate (i.e. normalized frequency) in the corpus of interest divided by that in the reference corpus. A step forward is the indication of statistical uncertainty in the form of confidence intervals for these ratios (cf. Gries 2022). This facilitates visual keyness assessments, by graphing ratios and their uncertainties. Further, negative binomial regression also supplies, for each item, a measure of dispersion for each corpus. This score reflects dispersion at the text level and can be translated into a directly interpretable index. Finally, and perhaps most importantly, the statistical assumptions of count regression models are in closer synchrony with the data: In contrast to the text-level techniques proposed in the literature, they are expressly designed to handle counts, i.e. non-negative integers. For illustration, I will use the BNC (Burnard 2007) to identify overrepresented verbs in academic writing. Limitations of the method will be given due consideration, and an online tutorial describes its application in R.

References

Bestgen, Y. 2014. Inadequacy of the chi-squared test to examine vocabulary differences between corpora. *Literary and Linguistic Computing* 29(2), 164–70.

Burnard*, L. 2007. Reference guide for the British National Corpus (XML Edition)*. http://www.natcorp.ox.ac.uk/XMLedition/URG/

Brezina, V. 2014. Effect sizes in corpus linguistics: keywords, collocations and diachronic comparison. Presented at the ICAME 2014 conference, University of Nottingham.

Brezina, V. & Meyerhoff, M. 2014. Significant or random? A critical review of sociolinguistic generalisations based on large corpora. *International Journal of Corpus Linguistics* 19(1), 1–28.

Cameron, A.C. & P.K. Trivedi. 2013. *Regression analysis of count data*. Cambridge: CUP.

Gries, S. Th. 2022. Toward more careful corpus statistics: uncertainty estimates for frequencies, dispersions, association measures, and more. *Research Methods in Applied Linguistics* 1(1), 100002.

Hardie, A. 2014. Log ratio – an informal introduction. http://cass.lancs.ac.uk/?p=1133

Kilgarriff, A. 2005. Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory* 1(2), 263–76.
Koplenig, A. 2019. Against statistical significance testing in corpus linguistics. *Corpus Linguistics and Linguistic Theory* 15(2), 321-346.
Lijffijt, J., Nevalainen, T., Säily, T., Papapetrou, P., Puolamäki, K. & Mannila, H. 2016. Significance testing of word frequencies in corpora. *Literary and Linguistic Computing* 31(2), 374–97.

Seeing the wood for the trees: Recursive partitioning with marginal effects plots

Lukas Sönning (University of Bamberg) & Jason Grafmiller (University of Birmingham)
lukas.soenning@uni-bamberg.de; j.grafmiller@bham.ac.uk

In the 10 years since their first appearance in the methodological literature aimed at linguists (Tagliamonte & Baayen 2012), recursive partitioning techniques, e.g. conditional inference trees and random forests, have gained considerable ground in corpus data analysis (see, e.g. Szmrecsanyi et al. 2016). There are settings where these tools offer attractive advantages over competing methods, and they are set to become "a new normal" in corpus linguistics. The literature often foregrounds a number of particularly useful features of these methods (e.g. Strobl et al. 2009; Levshina 2021): (i) tree-based methods are able to detect complex and non-linear relationships that may hold between (multiple) predictor variables and the outcome of interest; (ii) they work well even with small and/or imbalanced datasets; and (iii) it is often claimed that their primary outputs, i.e. decision-tree-like representations and variable importance scores, allow for straightforward interpretation (e.g. Szmrecsanyi et al. 2016).

In this paper, we argue that while "tree and forest" techniques may offer advantages over other methods, the current standard for the reporting of such analyses fails to capitalize on their full potential. Showing a single "best" tree and/or set of variable importance scores does not always yield easily interpretable data summaries. This is particularly true for complex settings, where these tools are meant to excel (Grafmiller, to appear). In fact, patterns are often quite difficult to read from individual trees, and interactions and non-linear effects, if they exist, are not at all apparent in the standard variable importance scores (Gries 2020). Recent work in machine learning has made some progress towards rendering the output of "black-box" modeling techniques more interpretable (e.g. Molnar 2019), and here we extend one of the key strategies—partial dependence plots (Friedman 2001)—for bringing into view the patterns suggested by a tree-and-forest-based analysis.

We incorporate techniques often found in the reporting of regression model analyses, more specifically, the kinds of model queries that produce marginal (or partial) effects plots. This involves the computation and summary of model-based predictions via purposeful manipulation of predictor values, which includes fixing certain inputs to specified values, to extract the equivalent of main effects and interaction effects. This allows us to assess the degree of complexity (or lack thereof) suggested by our analysis. Further, by tapping into the ensemble of trees constituting a random forest, we are able to offer indications of statistical uncertainty, similar to confidence bands or intervals around regression-based predictions. Translating random forests into focused marginal effects plots offers a clearer picture of patterns in the data and may also serve as an adjunct to regression analysis, pointing the analyst to (potentially) oversimplifying aspects of a standard linear model. We illustrate this process using natural data on the English genitive alternation (Grafmiller 2014).

References

Friedman, J. H. 2001. Greedy function approximation: A gradient boosting machine. *Annals of Statistics* 29(5), 1189-1232.

Grafmiller, J. 2014. Variation in English genitives across modality and genres. *English Language and Linguistics* 18(3), 471–496.

Grafmiller, J. To appear. Visualizing grammatical similarities in comparative variationist analysis. In L. Sönning & O. Schützler (eds.) *Data visualization in corpus linguistics: Reflections and future directions*, VARIENG.

Gries, S. Th. 2020. On classification trees and random forests in corpus linguistics: Some words of caution and suggestions for improvement. *Corpus Linguistics and Linguistic Theory* 16(3), 617-647.

Levshina, N. 2021. Conditional inference trees and random forests. In M. Paquot & S. Th. Gries, eds. *A practical handbook of corpus linguistics*. New York: Springer. 611-643.

Molnar, C. 2019. *Interpretable machine learning. A Guide for Making Black Box Models Explainable*. https://christophm.github.io/interpretable-ml-book/.

Strobl, C., Malley, J., & Tutz, G. 2009. An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods* 14(4), 323–348.

Szmrecsanyi, B., J. Grafmiller, B. Heller & M. Röthlisberger. 2016. Around the world in three alternations: Modeling syntactic variation in varieties of English. English World-Wide 37(2),109–137.

Tagliamonte, S. & R. H. Baayen. 2012. Models, forests and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change* 24(2), 135–178.

Modeling 21st-century Bangladeshi English through corpus data

Cristina Suárez-Gómez
cristina.suarez@uib.es

This presentation sets out to explore the development of English in Bangladesh and its present-day sociolinguistic situation, with a focus on its current phase of development and the linguistic relationship it has with other South-East Asian varieties (SAEs). English in Bangladesh has traditionally been considered a second language (L2), relegated to use in international communication, with Bengali as the national language, used by the majority of the population. Unlike some other varieties in the region, Bangladeshi English (BdE) has seen a notable revival in the 21st century, driven largely by globalization. This is reflected in a number of studies which typically consider BdE as part of a larger area of Englishes known as South-Asian Englishes (SAEs). These include consolidated varieties such as Indian English, Sri Lankan English and Pakistani English, as well as other, less well-known varieties such as Bhutan and Maldives Englishes. Individual linguistic studies of BdE are yet to emerge in the literature on World Englishes.

Using current models of the classification of English, specifically the *Dynamic Model of Postcolonial Englishes* (Schneider 2003, 2007) and the more recent *Extra- and Intra-territorial Forces Model* (Buschfeld and Kautzsch 2017, 2020), our first aim will be to account for the development of this variety, from its introduction into the territory in the 17th century to the present day. Although there is general consensus that BdE has reached the phase of 'nativization', BdE has followed a non-prototypical evolution in comparison to other Postcolonial Englishes. For this reason, we will seek to confirm BdE's 'nativization' with an analysis of data from GloWbE (Corpus of Web-Based Global English, Davies 2013), one of the few databases which contains language from the variety.

Second, this analysis also aims to explore linguistic similarities with other SAEs, such as Indian English, Sri Lankan English and Pakistani English, working from the extended hypothesis that Indian English, the largest institutionalised second-language variety of English, is emerging as the epicenter for English use in South Asia (Hundt 2013), and as such might be seen as serving as a model for the neighboring varieties. We have selected a list of specific morphosyntactic features reported as 'pervasive or obligatory' (label A) or 'neither pervasive nor extremely rare' (label B) in these three varieties, as represented in the *Electronic World Atlas of Varieties of English* (eWAVE, Kortmann et al. 2020), and have checked their use in GloWbE.

Results show that SAEs are homogeneous in the use of some of the linguistic features, and this would confirm the status of IndE as an epicenter. However, for some other features the picture is less clear and results are more heterogeneous, which would suggest that the varieties also exhibit individual developmental paths, these motivated by (i) different paces of evolution, affected by different extralinguistic factors, and (ii) the influence of different substrates, such as the clear presence of Bengali in the context of BdE, unlike in the case of Indian, Sri Lankan or Pakistan English.

References

Buschfeld, Sarah & Alexander Kautzsch. 2017. Towards an Integrated Approach to Postcolonial and Non-postcolonial Englishes. *World Englishes* 36(1). 104–126.

Buschfeld, Sarah & Alexander Kautzsch. 2020. *Modelling World Englishes. A joint approach to postcolonial and non-postcolonial varieties.* Edinburgh: Edinburgh University Press.

Davies, Mark. (2013) *Corpus of Global Web-Based English: 1.9 billion Words from Speakers in 20 countries* (GloWbE). https://corpus.byu.edu/glowbe/ (10 December, 2021).
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.
Kortmann, Bernd & Lunkenheimer, Kerstin & Ehret, Katharina (eds.) 2020. *The Electronic World Atlas of Varieties of English*. Zenodo. http://ewave-atlas.org (10 December, 2021).
Schneider, Edgar W. 2003. The Dynamics of New Englishes: From Identity Construction to Dialect Birth. *Language in Society* 79(2). 233–281.
Schneider, Edgar W. 2007. *Postcolonial English. Varieties around the world*. Cambridge: Cambridge University Press.

Epistemic verb expressions in native and non-native writing: Task effects (work in progress)

Daisuke Suzuki (UCL)
daisuke.suzuki.19@ucl.ac.uk

This study investigates the use of epistemic verb expressions (e.g. *I think*) in the written English of Japanese learners, compared to their use by British English native speakers. Much attention has been paid to the field of epistemic modality (Aijmer, 1997; Fung & Carter, 2007; Gablasova, Brezina, McEnery, & Boyd, 2017; Kaltenböck, 2010; Zhang & Sabet, 2016). Among other epistemic forms, the verb phrase *I think* is most frequently used. It acts as a pragmatic particle, which performs both tentative and deliberative functions as a hedge or booster (Holmes, 1990), e.g. "Perhaps she doesn't want to come, I think (a hedge)" (Zhang & Sabet, 2016, p. 335), vs. "I think that's absolutely right" (a booster) (Holmes, 1990, p. 187).

The use of *I think* has been studied from many perspectives. One example is the difference between English as a first language (L1) and second language (L2). Zhang and Sabet (2016) examined *I think* in the classroom across spoken data from L1 American English and L2 English of Chinese and Persian speakers. Their findings show that L2 groups use more *I think* than L1 groups and the positions of *I think* vary across the groups. Other literature focuses on the task effects. Gablasova et al. (2017) argued that higher proficiency L2 English learners could judge the use of the epistemic form considering the contexts. For example, they tend to use epistemic forms including *I think* more frequently in interactive tasks, which require them to adjust their (un)certainty towards their proposition such as a discussion in a language test setting, than monologic tasks such as a presentation.

In this study, the primary focus is on the task effects on the use of the epistemic verb expressions by Japanese learners of English (JLE); more specifically, whether tasks such as narrative or descriptive tasks affect their use of epistemic stance markers. Ongoing data analysis will be conducted on whether JLE adapt their writing style to reflect the contextual demands of the writing tasks to which they respond by employing epistemic verb expressions, applying the framework of systemic functional linguistics (Halliday, 1994). In addition, the current study examines the change of variation of epistemic verb expressions such as *I think* vs. others as JLE's proficiency level develops.

The L2 data are extracted from the EF Cambridge Open Language Database (EFCAMDAT). It contains 1.6 million words written by 3,441 JLE, whose proficiency levels range from A1 to C2 of the Common European Framework of Reference for Languages. The tasks the participants responded to cover a range of subjects such as advising a colleague or apologising to a client professionally. These speech acts (i.e. advice or apology) expect the writers to use epistemic modality as a hedge or booster. Since it is a learner corpus, L1 reference data have been collected specifically for this study using the same task descriptions and prompts to ensure that the analysis will be based on comparable data.

References

Aijmer, K. (1997). Modality in Germanic Languages. In S. Toril & J. W. Olaf (Eds.), *I think — an English modal particle* (pp. 1-48): De Gruyter Mouton.

Fung, L., & Carter, R. (2007). Discourse markers and spoken English: Native and learner use in pedagogic settings. *Applied Linguistics, 28*(3), 410-439.

Gablasova, D., Brezina, V., McEnery, T., & Boyd, E. (2017). Epistemic stance in spoken L2 English: The effect of task and speaker style. *Applied Linguistics, 38*(5), 613-637.
Halliday, M. A. K. (1994). An Introduction to Functional Grammar. In (2nd ed.): London: Arnold.
Holmes, J. (1990). Hedges and boosters in women's and men's speech. *Language & Communication, 10*(3), 185-205.
Kaltenböck, G. (2010). Pragmatic functions of parenthetical I think. *New approaches to hedging*, 243-272.
Zhang, G. Q., & Sabet, P. G. (2016). Elastic 'I think': Stretching over L1 and L2. *Applied Linguistics, 37*(3), 334-353.

Americanization in individual Finnish lingua franca English user's networks. Work in progress

Irene Taipale (University of Eastern Finland)
irene.taipale@uef.fi

This WiP paper reports my ongoing doctoral work that investigates individual lingua franca English users' sensitivity to ongoing language change in social networks. The study of variation and change in present-day Englishes has mostly focused on native language corpora aiming to minimize the role of an individual. It is only recently that individual variation in ongoing linguistic change has received serious scholarly attention (e.g. Anthonissen 2021), and non-native individuals have also been excluded from these discussions apart from some recent works (e.g Vetchinnikova & Hiltunen 2020).

This study combines the precision of qualitative methods with the empirical power of corpora by investigating ongoing change in individual repertoires through large sets of social network data. It explores whether there is a connection between the users' network properties and the frequency of incoming variants. Ongoing change is operationalized as Americanization, a variationist approach in which American English (AmE) and (BrE) variants constitute a simplified dichotomy that can be effectively quantified with corpus methods. What makes lingua franca English a fruitful testbed for the dynamics of linguistic diffusion is the lack of endogenous variety pressure to adhere to a certain variant, leaving room for external influences.

The variables include both orthographical and grammatical ones that set AmE and BrE apart (e.g. -or | -our; -er | -re; have gotten | have got) and constructions where AmE appears to lead frequency shifts (e.g. take a look | have a look; V + -ing | V + to-infinitive). The variables are initially chosen based on previous observations about native corpora that represent text types that are different from tweets (e.g. Leech et al. 2009; Baker 2017). For this reason, the study first examines tweets sent by individuals from the UK and US to verify the variables.

The primary data are retrieved from the Nordic Tweet Stream (Laitinen et al. 2018), a real-time corpus freely available at https://cs.uef.fi/nts/. Unlike texts in traditional corpora, tweets form an intricate web of interactions that can be effectively detected using computational techniques and modelled for their properties (network size, structure, similarity and frequency of communication). The dataset consists of 100 Twitter networks in which English is used as the main language of communication. Each network revolves around one user, the ego. Utilizing metadata that provides reliable information about the location of the users, the egos are chosen to represent different areas of Finland: the countryside, mid-size towns and large cities. The data are collected in January 2022 and amount to c. 200 million tokens.

Preliminary results suggest that users are polarized with their spelling choices but that they appear to strongly adhere to ongoing grammatical changes. Based on recent network-related observations (Laitinen & Lundberg 2020), a working hypothesis is that the larger the individual's network, the less likely it is that network density plays a role in how individuals adopt incoming linguistic features. Zooming in to the level of individuals and their networks can also open new avenues for the utilization of social media data in corpus-based sociolinguistics.

References

Anthonissen, L. (2021). Individuality in Language Change. Berlin: De Gruyter Mouton.

Baker, P. (2017). American and British English: Divided by a common language? Cambridge: Cambridge University Press.

Granovetter, M. (1973). 'The strength of weak ties', American Journal of Sociology 78 (6), 1360–1380.

Laitinen, M., Lundberg, J., Levin, M., Martins, R. (2018). The Nordic tweet stream: A dynamic real-Time monitor corpus of big and rich language data. 3rd Conference on Digital Humanities in the Nordic Countries, DHN 2018; Helsinki; Finland; 7 March 2018 through 9 March 2018, 2084, 349–362.

Laitinen, M. & Lundberg, J. (2020). 'ELF, Language Change, and Social Networks: Evidence from Real-Time Social Media Data'. In A. Mauranen & S. Vetchinnikova (eds) Language Change: The Impact of English as a Lingua Franca. Cambridge: Cambridge University Press, 179–204.

Leech, G., Hundt, M., Mair, C. & Smith, N. (2009). Change in Contemporary English: A Grammatical Study. Cambridge: Cambridge University Press.

Vetchinnikova, S. & Hiltunen, T. (2020). 'ELF and language change at the individual level'. In A. Mauranen & S. Vetchinnikova (eds) Language Change: The impact of English as a Lingua Franca. Cambridge: Cambridge University Press, 205–233.

Purpose subordinators on the move: *so, so that, just so* network

Elnora ten Wolde (University of Graz), elnora.ten-wolde@uni-graz.at
Gunther Kaltenböck (University of Graz), gunther.kaltenboeck@uni-graz.at

In the last 200 years, *just so* has arisen as a subordinator of condition (equivalent to *as long as*) (1) and as a subordinator of purpose (2) in the mid-19th century. In present day English, it is primarily used for purpose (see Kaltenböck and ten Wolde 2021). This paper discusses the evolution of the subordinator *just so* juxtaposed against the development of the formally and functionally related subordinators, *so* and *so that*. More precisely it asks how and why *just so* developed when there already existed two similar and long-standing subordinators of purpose, *so* and *so that*. To answer this question this study examines the activities of *so* and *so that* in the time period as the *just so* subordinator rose to power.

(1) *Toss that bird in the chuck or eat it yourself,* ***just so*** *you get it outa my sight.* (COCA)
(2) *Could Robin come to visit,* ***just so*** *the old woman could see someone from the reservation again?* (COCA)

The study is corpus-based and draws on data from the *Corpus of Contemporary American English* and the *Corpus of Historical American English*. Because the datasets for *so* and *so that* are so large, random samples of 200 tokens will be taken in five-year time periods, from 1900 to 2019, and coded for function.

An initial study, with a smaller sample size with 10-year periods, has shown that *so that*, which has either purpose or result meaning (Verstraete 2007), has seen a decline in the last century (from 140 tokens pmw in the 1930s to 71.6 in the 2000s). *So* maintains its predominant result meaning (e.g. Schiffrin 1987), but develops into a discourse marker (e.g. Bolden 2009). Building on these findings from Kaltenböck and ten Wolde (2021), this larger study will either substantialize this hypothesis or bring greater clarify to the changes taking part in this network at this time period.

In order to take a systemic look at the interdependency of these constructions, in the final stage of the project, we will model the findings in terms of a construction grammar network (e.g. Torrent 2015; Diessel 2019; Sommerer & Smirnova 2020), where these subordinators are either in competition or alternations on the micro-construction level (Cappelle 2006; Zehnetner 2019; Zehnetner & Traugott 2020). Ultimately, we hypothesize that *just so* fills the informal purpose niche, thus, bringing greater clarity to the fluctuating semantics of this family of constructions.

References

Bolden, Galina B. 2009. Implementing incipient actions: The discourse marker 'so' in English conversation. *Journal of Pragmatics* 41, 974-998.

Diessel, Holger. 2019. *The grammar network. How linguistic structure is shaped by language use*. Cambridge: Cambridge University Press.

Cappelle, B. 2006. 'Particle placement and the case for 'allostructions'.' In: D. Schönefeld (ed.), *Constructions all over: case studies and theoretical implications*. [Special issue of *Constructions*].

Kaltenböck, Gunther; Ten Wolde, Elnora. 2021. A *just so* story: On the recent development of the complex subordinator *just so*. In: Beatrix Busse, Nina Dumrukcic, & Ruth Möhlig-Falke (eds.), *Language and linguistics in a complex world. Data,*

interdisciplinarity, transfer and the next generation. Köln: University and City Library Cologne, pp. 114-118.
Schiffrin, Deborah. 1987. *Discourse Markers*. Cambridge: Cambridge University Press.
Sommerer, Lotte & Elena Smirnova (eds.). 2020. *Nodes and networks in diachronic construction grammar*. Amsterdam: John Benjamins.
Torrent, Tiago Timponi 2015. On the relation between inheritance and change: The constructional convergence and the Construction Network Reconfiguration Hypotheses. In Jóhanna Barðdal, Elena Smirnova, Lotte Sommerer and Spike Gildea (eds.), *Diachronic Construction Grammar*. Amsterdam/Philadelphia: John Benjamins, pp. 173–212.
Verstraete, Jean-Christophe. 2007. *Rethinking the coordinate-subordinate dichotomy: interpersonal grammar and the analysis of adverbial clauses in English*. Berlin: De Gruyter**.**
Zehentner, Eva. 2019. *Competition in language change: The rise of the English dative alternation*. Berlin & Boston: De Gruyter Mouton.
Zehentner, Eva; Traugott, Elizabeth Closs. 2020. Constructional networks and the development of benefactive ditransitives in English. In Lotte Sommerer and Elena Smirnova (eds.), *Nodes and networks in diachronic construction grammar*. Amsterdam: John Benjamins, pp. 167–212.

Should've did a corpus study, could've wrote a paper

Joe Trotta (Gothenburg University) & Monika Mondor (Gothenburg University)
joe.trotta@sprak.gu.se; monika.mondor@sprak.gu.se

An oft-noted feature of so-called non-standard dialects of English is the use of past tense forms instead of past participle forms in cases when those forms would be different, as in *did/done*, *wrote/written*, *saw/seen*, etc. (e.g., Anderwald, 2008). As noted in the literature, this typically occurs in connection with the perfective aspect as shown in the following examples (all attested in the COCA corpus):

i. *… but Stella had drove the mail to our house for years...*
ii. *…the hurt you feel for what you have did is worth the feeling of it…*
iii. *You couldn't have wrote it better…*

The use of the past tense form where a participle is required (referred to here for convenience as 'leveled participle' or 'LP') in these constructions is commonly noted as a mostly spoken alternative in many non-standard English varieties, in particular African-American Vernacular English (AAVE) or Southern White English (SWE) (cf. Munn and Tortora, 2014, Kortmann, 2006: 607, among others).

We see the LPs in these constructions as being more complex and nuanced than the previous research would indicate. In the present paper, we examine LPs through a systemic corpus study of the Corpus of Contemporary English (COCA) and the British National Corpus (the BNC), supplemented by additional attested data such as transcripts from podcasts, *YouTube* videos and online video games. Our aim is to shed critical light on what factors are connected to this variant as regards collocational, textual and situational factors as well as whatever regional or social variables may be involved. We show that, though the examples evidenced in the corpora are primarily spoken, LPs also occur in written texts and are not as limited to non-standard usages as the literature indicates.

In addition, we present data on this phenomenon outside of the perfective aspect (please note that for the purposes of this study we consider aspect and tense to be separate, but related, grammatical categories). Though it clearly occurs to a much lesser extent in other contexts, it can be found in passive constructions (e.g., *I am an American, and I feel what was did was wrong*). With a solid grounding in the corpus material, we present a richer picture of LP usage, we scrutinize previous analyses of the construction, give a brief view of this participial variant over time, and discuss whether the data indicates that the LP form is on the rise in Present-day English.

References

Anderwald, Lieselotte. 2008. The varieties of English spoken in the Southeast of England: morphology and syntax, in Kortmann, B. & C. Upton (eds), *Varieties of English. Vol 1: The British Isles*. Berlin and New York: Mouton de Gruyter, 440–462.

Kortmann, B. 2006. Syntactic variation in English: a global perspective. In Aarts, B. & A. McMahon (Eds.) *Handbook of English Linguistics*. Oxford: Blackwell. 603-24.

Munn, A. & C. Tortora. 2014. *Towards a Theory of Variation in English Participial Verb Forms: the Relevance of Auxiliary Morpho-Syntax*. ms. Michigan State University and CUNY.

Big Thyme for Big Data Methods - Approaching the Question of Homophone Durations with a Large Automatically Annotated Dataset

Peter Uhrig (Friedrich-Alexander-Universität Erlangen-Nürnberg)
peter.uhrig@fau.de

In her seminal study *Time and Thyme are not Homophones: The Effect of Lemma Frequency on Word Durations in Spontaneous Speech*, Gahl (2008) divided pairs of homophonous words into a high-frequency and a low-frequency group and was able to show that the high-frequency items were on average shorter than the low-frequency items, confirming her hypothesis that lemma frequency plays a role and thus that phonetic production cannot operate on a simple representation of words as strings of phonemes or speech sounds in our mental storage. In a re-analysis of Gahl's work, Lohmann (2018) found that although Gahl's statistics were problematic, her results hold and her conclusions are thus supported by the data (and by another dataset discussed in Lohmann 2017).

It must be noted, however, that the word *thyme* is quite rare. As noted by Lohmann (2018: 183) himself, the word occurs exactly once in the Switchboard corpus used in both studies, where it is in fact used as part of the compound *lemon thyme*. While the overall number of tokens in the dataset is perfectly good at roughly 80,000, the distribution is extremely skewed, with a few high-frequency items (such as *time* at 7,312 according to Lohmann) accounting for the bulk of that number.

The NewsScape English Corpus (Uhrig 2018) on the other hand contains more than 1,000 occurrences of the word *thyme* alone in 2 billion running words, although sometimes it is of course not entirely clear what counts as *time* and *thyme*, as in the example video found behind the QR code on the right (click or scan the code - see Uhrig 2020 for a description of permalinks with QR codes in the Red Hen ecosystem). The corpus is based on the UCLA Library Broadcast NewsScape, which contains recordings of American TV news (in a broad sense) together with text files containing the subtitles. All files were processed in an NLP pipeline and run through forced alignment software, which also provides the length of the word.

In a first naïve pilot study based on a small set of pairs and a maximum of 10,000 hits per item, the differences observed were often very small and, most interestingly, did not always show that the more frequent item in the corpus was the shorter word. Thus, for instance, the average instance of *steak* is significantly longer than the average instance of *stake*, even though the latter is roughly three times more frequent in our corpus and roughly 5 times more frequent in the Corpus of American Soap Operas (Davies 2011-), which is meant to represent informal language.

In this presentation, I will offer a range of statistics and discuss whether results as general as those by Gahl (a) can be replicated with big data methods, and (b) how much variation between lexical items exists.

References

Davies, Mark (2011-) *Corpus of American Soap Operas*. Available online at https://www.english-corpora.org/soap/.

Gahl, Susanne (2008) "*Time* and *Thyme* are not Homophones: The Effect of Lemma Frequency on Word Durations in Spontaneous Speech." *Language* 84/3, 474-496.

Lohmann, Arne (2018) "*Time* and *thyme* are NOT homophones: A closer look at Gahl's work on the lemma-frequency effect, including a reanalysis." *Language* 94/2, e180-e190.

Uhrig, Peter (2018) "NewsScape and the Distributed Little Red Hen Lab – A digital infrastructure for the large-scale analysis of TV broadcasts." In: Anne-Julia Zwierlein, Jochen Petzold, Katharina Böhm and Martin Decker (eds.), *Anglistentag 2017 in Regensburg: Proceedings. Proceedings of the Conference of the German Association of University Teachers of English.* Trier: Wissenschaftlicher Verlag Trier.

Uhrig, Peter (2020) "Multimodal Research in Linguistics." *ZAA* 68/4, 345-349.

The adaptation of a corpus: Reformatting CANBEC for sociolinguistic analysis

Isolde van Dorst (Vienna University of Economics and Business)
isolde.van.dorst@wu.ac.at

The *Cambridge and Nottingham Business English Corpus* (CANBEC) (Handford, 2010) is unique in its size, style, composition, and metadata. It is a one million word corpus of spoken business English that consists of 64 different meetings between 361 speakers. The metadata that is part of this corpus pertains information regarding the speakers, the company, and the conversation at hand, totalling information on 20 different variables across these three entities. This corpus, due to its unique and valuable content to other businesses worldwide, is not publicly available, nor was it ever intended for public release. Its compiler, Michael Handford, as well as a handful of other people, have been able to conduct research on spoken business communication using the data that was collected, in a mainly qualitative manner. The format of the original CANBEC corpus consists of the transcribed conversations in text files with additionally transcribed conversational information. The metadata is contained in a single spreadsheet, in which metadata from the speakers, the companies, and the conversation itself are all mixed together. While this format has proven to work for the analysis of spoken English business communication overall through discourse analytical methods, it does not provide a clear way of analysing social variables quantitatively or on a larger scale beyond individually chosen speakers.

As part of my PhD, I have converted the CANBEC corpus from its original format into a CQPweb-compatible (Hardie, 2012) database, which allows for computational, sociolinguistic analyses that would originally have been quite time-consuming to produce. The goal of this endeavour was to create a computational connection between the transcribed conversations and the metadata files, all while maintaining the high level of detail that was originally included in both the transcription of the meetings as well as the metadata itself. I have chosen to convert to a CQPweb-compatible format, as CQPweb allows for many sociolinguistic analyses based on the metadata in the corpus. Ultimately, this reconfiguration of CANBEC means that it is now possible to conduct more quantitative, statistical analyses on a very rich corpus that will provide new/more information on spoken English business communication.

This reformat of CANBEC draws upon lessons learnt from the compilation of the *British National Corpus 2014* (Love et al., 2017), which is a similarly rich corpus in terms of metadata, though it needed adaptation for the business context of CANBEC and its combination of conversational information and thorough metadata. During this talk, I hope to give an overview of the different steps I had to undertake to complete this reformatted dataset, the problems that I ran into along the way, and how I handled those to stay as closely to my goal of maintaining the level of detail from the original CANBEC corpus as possible. Additionally, a short sociolinguistic case study of the politeness marker "thank you" will show how this adaptation of CANBEC contributes to furthering our understanding of business communication.

References

Handford, M. (2010). *The Language of Business Meetings*. Cambridge University Press. https://doi.org/10.1017/cbo9781139525329

Hardie, A. (2012). CQPweb — combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, *17*(3), 380–409.

https://doi.org/10.1075/ijcl.17.3.04har
Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken BNC2014: designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics*, *22*(3), 319–344. https://doi.org/10.1075/ijcl.22.3.02lov

Analysing category change with enriched data: A grammatical and sociolinguistic study of *-ed* participles, 1810–2009

Turo Vartiainen & Tanja Säily (University of Helsinki)
tanja.saily@helsinki.fi

This paper focuses on a relatively recent change affecting *-ed* participles (e.g., *interested*, *amazed*, *surprised*). According to previous research, these participles have become more adjective-like in recent history through a change in their degree modification patterns: until the early nineteenth century, the participles were typically modified with *much* (e.g., *he's much interested in it*), while in Present-day English, the preferred modifier is *very* (e.g., *he's very interested in it*; Denison 1998; Vartiainen 2021). Earlier studies have shown that the change took place gradually from the mid-nineteenth century to the mid-twentieth century, but many details pertaining to the precise path of change remain to be explored both from a grammatical and a sociolinguistic perspective. Moreover, the datasets used have been small, and the analyses have focused on token frequency rather than the productivity or type frequency of the patterns, which means that the figures could have been skewed by a few frequent participles.

We aim to fill this gap by investigating the 200-million-word fiction section of the *Corpus of Historical American English* (COHA; Davies 2010–). Our first goal is to examine the grammatical context of the change more closely and to explore the possibility that the shift from *much* to *very* may have been facilitated by an intermediate stage, where *much* was modified by *very* (e.g., *she was very much frightened by it*). In other words, we explore the possibility that the path of development proceeded as follows: *much* → *very much* → *very*. We also zoom in on a class of *-ed* participles denoting psychological states, a semantic property that has previously been argued to be central to the change. Our second goal is to investigate the change from a sociolinguistic perspective in order to see whether the change was led by men or women, bearing in mind the observations in Nevalainen and Raumolin-Brunberg (2003), who found that many grammatical changes in the history of English have been spearheaded by women. Our sociolinguistic examination makes use of gender metadata developed by Öhman et al. (2019) to enrich the existing metadata of COHA; Öhman et al. were able to identify the gender of the author for c. 90% of the fiction texts. We use robust statistical methods that enable the diachronic comparison of competing patterns across subcorpora in terms of proportions of types (Rodríguez-Puente et al. in press; see Figure 1 for an example).

Our initial results support our hypothesis regarding the bridging context: the proportion of *very much -ed* types out of all *much -ed* types is the highest from c. 1860 to c. 1940, which overlaps temporally with the increased proportion of the *very -ed* pattern and the loss of *much* as an individual modifier. However, when it comes to gender, there are no statistically significant differences in the adoption of *very* between men and women; the change seems to have progressed at a relatively even pace for both genders (Figure 1). Nevertheless, from a methodological perspective, we argue that enriched datasets like the one used in our study show much promise for future research, and the practice of investigating existing corpora in light of new metadata may in fact become the "new normal" in corpus linguistics.

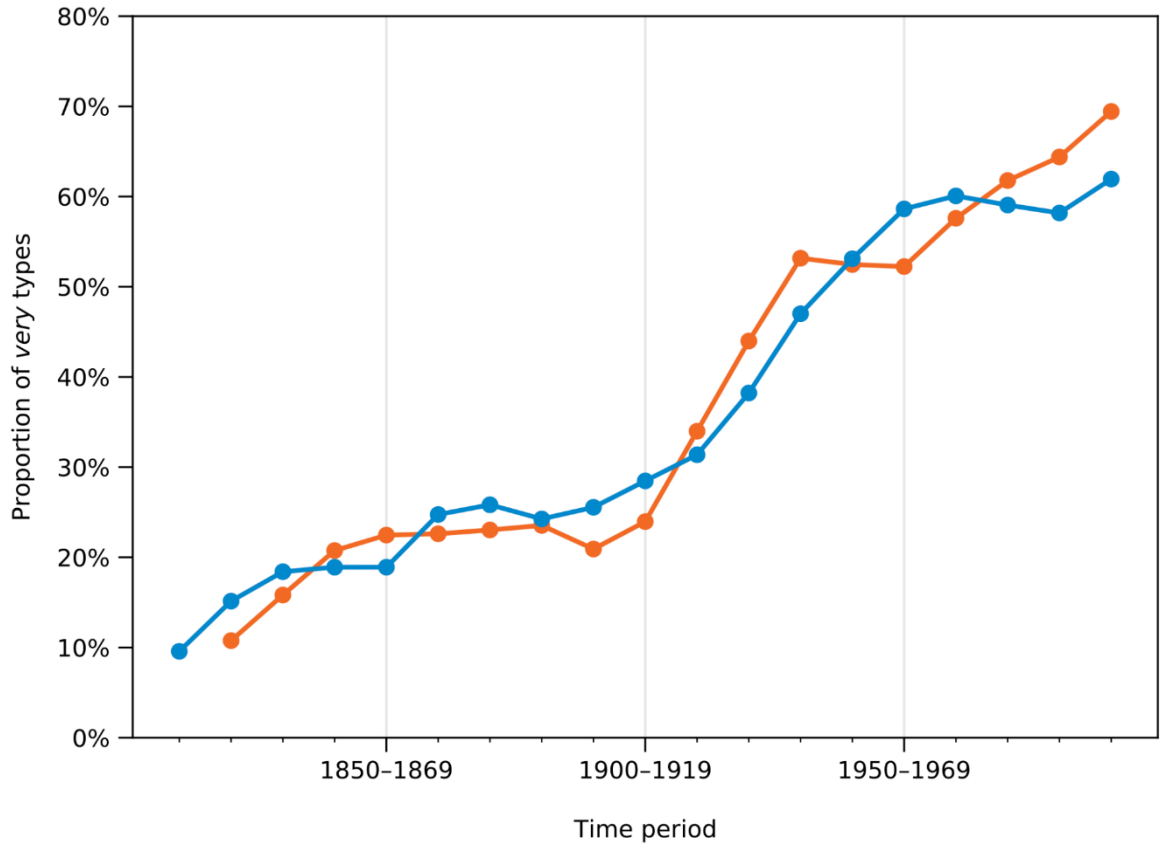

Figure 1. Proportion of *very -ed* types out of *very -ed* and *much -ed* types in the fiction section of COHA over time (blue = men, orange = women). 20-year sliding window, 10-year intervals. Curves: randomly sampled subcorpora with a sum of 100 *very -ed* and *much -ed* types.

References

Davies, Mark. 2010–. The Corpus of Historical American English: 400 million words, 1810–2009. https://www.english-corpora.org/coha/

Denison, David. 1998. Syntax. In Suzanne Romaine (ed.), *The Cambridge history of the English language, vol. 4*, 1776-1997, 92–329. Cambridge: Cambridge University Press.

Nevalainen, Terttu & Helena Raumolin-Brunberg. 2003. *Historical sociolinguistics: Language change in Tudor and Stuart England*. London: Pearson Education.

Öhman, Emily, Tanja Säily & Mikko Laitinen. 2019. Towards the inevitable demise of *everybody*? A multifactorial analysis of *-one/-body/-man* variation in indefinite pronouns in historical American English. Paper presented at the 40th Annual Conference of the International Computer Archive of Modern and Medieval English (ICAME 40), Neuchâtel, Switzerland, June 2019. https://tanjasaily.fi/talks/icame40_ohman_et_al_2019.pdf

Rodríguez-Puente, Paula, Tanja Säily & Jukka Suomela. In press. New methods for analysing diachronic suffix competition across registers: How *-ity* gained ground on *-ness* in Early Modern English. *International Journal of Corpus Linguistics*.

Vartiainen, Turo. 2021. Trends and recent change in the syntactic distribution of degree modifiers: Implications for a usage-based theory of word classes. *Journal of English Linguistics* 49(2): 228–251.

Chunking at the individual level

Svetlana Vetchinnikova (University of Helsinki)
svetlana.vetchinnikova@helsinki.fi

Usage-based accounts of language hold that grammar emerges from usage through the application of domain-general cognitive processes to concrete instances. Chunking is one such domain-general cognitive process which is postulated to play a key role in the formation of new units. With repeated usage elements of an expression become chunked, undergo reanalysis and acquire a new meaning. A well-known example is *I don't know,* which is phonologically reduced and conveys an additional pragmatic function of mitigated disagreement when used as a unit (Bybee & Scheibman 1999). Reduction is in general proposed as a diagnostic of a change in the internal structure of an expression. Normally, reanalysis is observed at the communal level of language representation, that is in data aggregated across different speakers of a given language community. Does it operate at the individual level, within the language use of a single speaker? And more specifically, can we use reduction to identify constructions which are undergoing reanalysis in one's idiolect?

This paper will harness the variation between contracted and uncontracted forms of *it is* to identify chunks which are possibly undergoing reanalysis within an idiolect. *It is/it's* occurs in a variety of syntactic structures: clefts, progressives, passives, extraposed and copular structures, but the range of constructions it participates in is even wider making it an interesting case. In this paper, I will rely on COBUILD Grammar Patterns (Francis et al. 1996, 1998) to categorize the constructions (see Hunston 2019; Perek & Patten 2019 for a discussion of the connection between grammar patterns and constructions). Chunks will be operationalized as lexically specified constructions. In principle, if a chunk is associated with a reduced form, it is possible that it is undergoing a structural change. Previous research has identified multiple factors which can have an effect on the variation between contracted and uncontracted forms (Labov 1969; MacKenzie 2012; Barth & Kapatsinski 2017; Mair 2017; Vetchinnikova & Hiltunen 2020). These factors will be controlled for.

As my data I will use several longitudinal corpora of comments posted on a single blog by different individuals over 8 years. The corpora vary in size between 1.75 million and 160 thousand words. The largest corpus contains 10 thousand occurrences of *it is/it's*. The comments of over 4 thousand occasional commenters (ca. 3.5 million words in total) on the same blog can serve as a reference corpus representing the communal level. For each corpus, I will build a logistic regression model predicting the contracted form and include the following independent variables as fixed effects: priming (primed/not primed), construction (e.g. *it is/it's* ADJ that, *it is/it's* det N to-inf), construction token frequency, construction type frequency, relative frequency of a lexical item filling the open slot, frequency of a lexical item immediately after *it is/it's*, as well as order of occurrence to test for change over time. To examine whether some lexically specified constructions (or in other words chunks) show a stronger association with the reduced form, I will enter lexical items filling the open slots as random effects. The model will be implemented in R using the lme4 package (Bates et al. 2015).

References

Barth, Danielle & Vsevolod Kapatsinski. 2017. A multimodel inference approach to categorical variant choice: Construction, priming and frequency effects on the choice

between full and contracted forms of *am, are* and *is*. *Corpus Linguistics and Linguistic Theory* 13(2). 203–260.
Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67. 1–48.
Bybee, Joan & Joanne Scheibman. 1999. The effect of usage on degrees of constituency: The reduction of *don't* in English. *Linguistics* 37(4). 575–596.
Francis, Gill, Susan Hunston & Elizabeth Manning. 1996. *Collins COBUILD Grammar Patterns 1: Verbs*. London: HarperCollins.
Francis, Gill, Susan Hunston & Elizabeth Manning. 1998. *Collins COBUILD Grammar Patterns 2: Nouns and Adjectives*. London: HarperCollins.
Hunston, Susan. 2019. Patterns, constructions, and applied linguistics. *International Journal of Corpus Linguistics* 24(3). 324–353.
Labov, William. 1969. Contraction, deletion, and inherent variability of the English copula. *Language* 45(4). 715–762.
MacKenzie, Laurel E. 2012. Locating Variation Above the Phonology. PhD thesis. University of Pennsylvania.
Mair, Christian. 2017. From priming and processing to frequency effects and grammaticalization? Contracted semi-modals in present-day English. In Marianne Hundt, Sandra Mollin & Simone E. Pfenninger (eds.), *The Changing English Language*, 191–212. Cambridge: Cambridge University Press.
Perek, Florent & Amanda L. Patten. 2019. Towards an English Constructicon using patterns and frames. *International Journal of Corpus Linguistics* 24(3). 354–384.
Vetchinnikova, Svetlana & Turo Hiltunen. 2020. ELF and language change at the individual level. In Anna Mauranen & Svetlana Vetchinnikova (eds.), *Language Change*, 205–233. Cambridge University Press.

On the importance of the common ground in scientific discourse: extenders andfocus operators in late Modern English

Iria Bello Viruega & Estefanía Sánchez Barreiro (Universidade da Coruña)
fannillasb@gmail.com

To be relevant and coherent, speakers include guidelines in their messages to trigger mental representations based on inferences (Sanders & Pander Maat, 2006). Extension and focalization are key strategies to structure information in discourse. Speakers can limit the potential ambiguity of discourse to guide inferential processes and maximize cognitive effort in processing (Blakemore 1987, Sperber & Wilson 1995, 2002, Levinson 2000, Wilson & Sperber 2002). In this study, we contribute to the long-standing debates about the semantic and pragmatic nature of scales by investigating two structures that activate two opposite inferencing pathways, namely extenders (*and others*, *and things like that*…) and focus operators (*also, even*…) in late Modern English scientific register. Extenders have traditionally been considered paradigmatic examples of vague language (Channel 1994, Williamson 1994, Jucker, Smith & Lüdge 2003, Drave 2004, Cutting 2007, Sánchez Barreiro 2020). Conversely, focus operators have been classified as effective devices that help restrict the number of inferences needed to understand an utterance (König 1991, Dimroth & Klein 1996, Traugott 1996, Gast 2006, 2017a, 2017b, Traugott 2006, Gast & van der Auwera 2011, Gast & Rzymski 2015, Loureda et al. 2015, Nadal, Recio Fernández, Rudka, & Loureda 2017, Cruz & Loureda 2019). We argue that these two structures are in fact very similar cognitively, as they both generate similar categorization processes that allow receivers to create webs of coherence and understand what is communicated in relation to previous discourse as well as to their world knowledge.

The corpus material for this study was taken from two of the subcorpora of the *Coruña Corpus of English Scientific Writing* (Moskowich & Crespo García 2007, Moskowich & Parapar López 2008, Crespo García & Moskowich 2010, 2020), namely history (*CHET*, Moskowich, Lareo, Lojo Sandino & Sánchez Barreiro 2019) and life sciences (*CELiST*, Lareo, Monaco, Esteve-Ramos, Moskowich, 2020). Each subcorpus contains two texts per decade written by English-speaking authors (800,000 analyzable words in total). The corpus covers the 18th and 19th centuries. The diachronic variable was explored to account for differences in the use of extenders and focus operators in the period of consolidation of the scientific register in English. Additionally, the diaphasic dimension is studied to untangle differences in texts aimed at specialized and non-specialized audiences.

Results showed a correlation between the level of specialization of texts and frequency of use. The trends outlined after data analysis provide a new dimension not yet explored in the study of extenders and focus operators. They should be taken as a starting point for future research that could result in a more accurate description of the English scientific register in the 18th and 19th centuries.

References

Blakemore, Diane. (1987). *Semantic constraints on relevance.* Oxford: Blackwell.
Channell, J. (1994) (ed.) *Vague Language*. Oxford: Oxford University Press.
Cutting, J. (2007) (ed.) *Vague Language Explored*. Basingstoke: Palgrave Macmillian.
Crespo García, B. & Moskowich, I. (2010). CETA in the Context of the Coruña Corpus. *Literary and Linguistic Computing, 25*(2): 153–164.

Crespo García, B. & Moskowich, I. (2020). Astronomy, Philosophy, Life Sciences and History Texts: Setting the Scene for the Study of Modern Scientific Writing. *English Studies*.

Cruz, A. & Loureda, Ó. (2019). Processing patterns of focusing in Spanish. En Ó. Loureda, I. Recio Fernández, L. Nadal & A. Cruz (Eds.), *Empirical Studies of the Construction of Discourse*. Amsterdam: Benjamins.

Dimroth, Ch. & Klein, W. (1996). Fokuspartikeln in Lernervarietäten. Ein Analyserahmen und einige Beispiele. *Zeitschrift für Literaturwissenschaft und Linguistik 104*, 73–114.

Drave, N. (2002) Vaguely speaking: a corpus approach to vague language in intercultural conversations, in P. Peters, P. Collins and A. Smith (ed.), *New Frontiers of Corpus Research. Papers from the Twenty First International Conference on English Language Research on Computerized Corpora*. Sydney, 2000. Amsterdam & New York: Rodopi. Available on-line from http://www.ingentaconnect.com/content/rodopi/lang/2001/00000036/00000001/art00003

Gast, V. (2006). The Distribution of *also* and *too*: a Preliminary Corpus Study. *Zeitschrift für Anglistik und Amerikanistik* 54(2), 163-176.

Gast, V. (2017a). 'So much as' and 'even' in downward entailing contexts. A quantitative study based on data from the British National Corpus.

Gast, V. (2017b). The scalar operator even and its German equivalents: Pragmatic and syntactic factors determining the use of *auch*, *selbst* and *sogar* in the Europarl corpus. En De Cesare, A.M. & Andorno, C. (Eds.), *Focus on additivity. adverbial modifiers in Romance, Germanic and Slavic languages*. Amsterdam/Philadelphia: John Benjamins, 201-234.

Gast, V. & van der Auwera, J. (2011). Scalar Additive operators in the Languages of Europe. *Language* 87(1): 2-54.

Gast, Volker, & Christoph Rzymski. (2015). Towards a corpus-based analysis of evaluative scales associated with *even*. *Linguistik Online, 71*/2.

Jucker, Andreas, Sara W. Smith & Tanja Lüdge (2003). Interactive aspects of vagueness in conversation. *Journal of Pragmatics 35*(12), 1737-1769.

König, Ekkehard. (1991). *The Meaning of focus particles: a comparative perspective.* London: Routledge. https://doi.org/10.4324/9780203212288.

Lareo, Inés, Monaco, Leida M., Esteve-Ramos, María José & Moskowich, Isabel (comps.). 2021. *Corpus of English Life Sciences Texts*. A Coruña: Universidade da Coruña.

Levinson, Stephen. (2000). *Presumptive meanings – The theory of generalized conversational implicature.* Cambridge (MA): MIT Press.

Loureda, Ó., Cruz, A., Rudka, M., Nadal, L., Recio, I., & Borreguero, M. (2015). Focus Particles in Information Processing: An Experimental Study on Pragmatic Scales with Spanish incluso. Focus particles in the Romance and Germanic languages. Experimental and corpus-based approaches, *Linguistic online, 71*(2), 129-151.

Nadal, L., Recio Fernández, I., Rudka, M. & Loureda, Ó. (2017). Processing additivity in Spanish: incluso vs. además. En A. De Cesare & C. Andorno (Eds.), *Focus on Additivity. Adverbial modifiers in Romance, Germanic and Slavic languages.* Amsterdam: Benjamins, 137-154.

Moskowich, Isabel and Crespo, Begoña (2007). Presenting the Coruña Corpus: A Collection of Samples for the Historical Study of English Scientific Writing. In Pérez Guerra, Javier et al. (eds.) *'Of Varying Language and Opposing Creed': New Insights into Late Modern English*. Bern: Peter Lang. 341–357.

Moskowich, Isabel, Lareo, Inés, Lojo Sandino, Paula & Sánchez-Barreiro, Estefanía (comp.). 2019. *Corpus of History English texts*. A Coruña: Universidade da Coruña.

Moskowich, I. & Parapar López, J. (2008). Writing Science, Compiling Science. The Coruña Corpus of English Scientific Writing. In Lorenzo Modia, M.J. (Ed.), *Proceedings from the 31st AEDEAN Conference*, 531–544.

Sánchez Barreiro, Estefanía (2020). *Los* extenders *como recurso estilístico en el discurso científico inglés del siglo XVIII*. Doctoral dissertation, Universidade da Coruña.

Sanders, Ted & Henk Pander Maat. (2006). Cohesion and coherence: linguistic approaches. In Keith Brown (Ed.), *Encyclopedia of language and Linguistics* (591–595). London: Elsevier (Volume 2). https://doi.org/10.1016/B0-08-044854-2/00497-1

Sperber, Dan & Deirdre Wilson. (1995). *Relevance*. Oxford: Blackwell.

Sperber, Dan & Deirdre Wilson. (2002). *Pragmatics, modularity and mind-reading. Mind and Language, 17*, 3–23. https://doi.org/10.1111/1468-0017.00186

Traugott, E.C. (2006). The Semantic Development of Scalar Focus Modifiers. En van Kemenade, A. (Ed*.), The Handbook of the History of English*. Oxford: Blackwell, 335-359.

Williamson, Timothy, 1994. Vagueness. In: Asher, R., Simpson, J. (Eds.), *The Encyclopedia of Language and Linguistics*. Pergamon Press, Oxford, pp. 4869–4871.

Wilson, Deirdre & Dan Sperber. (2002). Relevance theory. Working papers in Linguistics, 14, 249-287.

Accurate confidence intervals on Binomial proportions, functions of proportions and related scores

Sean Wallis (Survey of English Usage - UCL)
s.wallis@ucl.ac.uk

In many fields, confidence intervals are growing in popularity, and they are increasingly becoming mandatory in journals. Plotting data and citing scores with confidence intervals can convey a model of the distribution of sampling uncertainty to the reader that is absent from traditional approaches where plotting data and conducting analysis are separated. Researchers may compare sampled scores for significant difference or compare them with a benchmark score.

However, many statistical sources employ a standard error formula that is frequently incorrect. This assumes that the probable true value of an observed parameter is Normally distributed, an assumption rarely true for small samples or observations near numeric bounds. This generates intervals that are not consistent with standard statistical test procedures, and occasionally, produce wholly implausible results.

In this paper we discuss a superior approach to constructing intervals for a wide range of properties. This builds on the Wilson score interval for the simple proportion p, which is robust on the probability scale P = [0, 1] and may be corrected for continuity and sampling. We demonstrate how we may compute intervals for properties that are functions of p (such as ln(p), logit(p) and p^2), and, by employing Zou and Donner's interval difference theorem, for algebraic combinations of independent proportions p1, p2, etc. (such as p2 – p1, Σpi, p1 / p2 and p1^p2). These methods are efficient to calculate, robust, and perform consistently with standard tests, while being capable of extension to novel statistical test procedures.

"Sitting *on* a chair writing a paper *on* prepositions": a cognitive semantic study of the polysemy of the preposition *on*

Michelle Weckermann (University of Augsburg)
michelle.weckermann@philhist.uni-augsburg.de

This paper examines the polysemy of the preposition *on* as part of a larger project investigating a range of prepositions. Using a cognitive semantic approach, the different senses of *on* are analysed and modelled with image schemas and semantic networks.

While there is extensive research on the polysemy of prepositions (e.g. Hanazaki, 2005 for *by*; Tyler & Evans, 2003 for *over*), many studies based their analyses on fabricated examples (e.g. Tyler & Evans, 2003; Lakoff, 1987). Moreover, many studies have been criticised for relying solely on the researchers' introspective judgments and thus lacking a methodology for determining and distinguishing senses of a preposition (Sandra & Rice, 1995). Tyler and Evans' (2003) principled polysemy approach to *over* was the first to propose a set of methodological criteria for determining an established sense and the central sense of a preposition.

The present study aims to improve on the two aforementioned pieces of criticism by drawing on natural data from corpora instead of fabricated examples, and by applying methodological criteria for determining and distinguishing senses. In relation to the first point, the data for *on* was gathered from a selection of corpora, including a legal corpus (EuroParl), as well as four novels from different genres (thriller, romance/drama, dystopia/fantasy, and philosophical novels). The corpora are representative of a range of genres and topic areas, which should ideally mirror as many of the different nuances of meaning manifested in the preposition's senses as possible and therefore increase the naturalness of the data.

Concerning the second point of criticism relating to the methodology employed for determining and distinguishing senses, Tyler and Evans' (2003) two criteria state that an established sense has to express a distinct nuance of meaning (e.g. a distinct spatial configuration) from the other, already existing senses, and has to do so independently of context. These two criteria are adopted in the present study yet specified with ideas from Cruse's (2000) account of how different types of contexts can influence word meaning, in order to ensure, for instance, that a sense is indeed context-independent and does not gain its meaning from contextual information or encyclopaedic knowledge (in line with the second criterion). Applying these criteria to *on,* it is argued, for instance, that the preposition does not have a sense denoting 'motion to a position on top of something', as this dynamic meaning is not provided by the preposition itself but rather by the preposition in combination with its immediate sentential context – a motion verb. The analysis of the entire corpus-based data set for *on* yielded a total of ten distinct senses of spatial, temporal and abstract nature.

References

Cruse, D. A. (2000). *Meaning in Language.* Oxford: Oxford University Press.

Hanazaki, M. M. (2005). Toward a model of principled polysemy. *English Linguistics,* 22(2), pp. 412-442.

Lakoff, G. (1987). *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind.* Chicago: The University of Chicago Press.

Sandra, D. & Rice, S. (1995). Network analyses of prepositional meaning: Mirroring whose mind – the linguist's or the language user's? *Cognitive Linguistics,* 6(1), pp. 89-130.

Tyler, A. & Evans, V. (2003). *The Semantics of English Prepositions.* Cambridge: Cambridge University Press.

Pragmatic Variation in World Englishes: A corpus-pragmatic analysis of question tag use in Nigerian, Philippine, and Trinidadian English

Michael Westphal (University of Münster)
michael.westphal@wwu.de

Corpus-based research on World Englishes has largely focused on morpho-syntax, lexico-grammar, and lexicon (e.g. Hundt & Gut 2012), while pragmatic phenomena were long neglected. The field of variational pragmatics (Schneider & Barron 2008) addresses this research gap but corpus-based studies on the pragmatics of New Englishes (Englishes learned as L2) are still rare. In addition, text type variation has not been addressed sufficiently in variational pragmatics (Barron 2015: 224), and most studies focus on individual forms rather than entire variant fields (e.g. Unuabonah & Oladipupo 2018).

This paper analyzes the variation in the use of question tags in Nigerian, Philippine, and Trinidadian English regarding form and pragmatic function. I analyze variant question tags (e.g. *do you*) as well as English (e.g. *right*) and indigenous (e.g. Yoruba *abi*) invariant question tags in six dialogue text types from the Nigerian, Filipino, and Trinidadian component of the International Corpus of English (ICE): conversations, phonecalls, broadcast dialogues and interviews, classroom lessons, and legal cross-examinations. I address the following research questions:

- Which question tag forms do Filipino, Nigerian, and Trinidadian speakers use?
- How does text type influence the overall distribution of question tag forms in the three varieties?
- How do text type and function constrain the selection of particular forms over others in the three varieties?

To identify all forms that may function as question tags, all 347 texts (~750,000 words) were read 'manually' (horizontally). The question tags identified in this process were coded qualitatively for their pragmatic function, distinguishing between informative, punctuational, and facilitative uses. I identified overall 4002 tokens across the three varieties, which were further analyzed using binary regression models with form (e.g. *right* vs. all other forms) as the dependent variable and variety, text type, and function as predictor variables.

In all three varieties question tags exhibit very similar frequencies (~5,300-5,700 tokens per million words) and there is a similar distribution across text types. The three New Englishes are characterized by a low frequency of variant question tags, a mix of many different English and indigenous invariant forms, and multilingual variation is constrained in similar ways by text type: indigenous forms are mostly used in private conversations but individual forms, such as Tagalog *'no* in Philippine English, may be used in more formal text types. The regression models show that there are similarities and differences in terms of text type variation and the form-function relationship of selected forms: for example, *OK* is a universal teacher tag used to integrate students into the classroom discourse; *right* is very frequent in Trinidadian English and has a very diverse usage profile, while its use is more restricted in Philippine and Nigerian English.

The analysis demonstrates a high degree of nativization at a pragmatic level as question tag use is highly multilingual in the three New Englishes. Text type is shown to be an essential factor to understand pragmatic variation in World Englishes. Finally, the paper illustrates the benefits of a variationist analysis of an entire field of variants rather than focusing on individual forms.

References

Barron, A. (2015). “And Your Wedding Is the Twenty-Second of June Is it?”. In C. P. Amador-Moreno (Ed.), *Pragmatic markers in Irish English* (pp. 203–228). Amsterdam: Benjamins.

Hundt, M. & U. Gut (Eds.) (2012). *Mapping unity and diversity world-wide: Corpus-based studies of New Englishes*. Amsterdam: Benjamins.

Kortmann, B. & E. Schneider (Eds.) (2008). *A handbook of varieties of English*. Berlin: de Gruyter.

Schneider, K. & A. Barron (Eds.) (2008). *Variational pragmatics*. Amsterdam: Benjamins.

Unuabonah, F. O. & R. O. Oladipupo (2018). “You’re Not Staying in Island Sha O”: O, Sha and Abi as Pragmatic markers in Nigerian English. Journal of Pragmatics 135, 8–23.

Discourses of 21st century identity documents in the UK: A contribution to diachronic corpus studies

Viola Wiegand
v.wiegand@bham.ac.uk

Going into the third decade of the 21st century, an increasing number of large-scale and diachronic corpora are becoming widely available. These developments are facilitating some research into discursive developments over time. Corpus linguists continue to explore best practices for identifying discourses in such large corpora – or resort to compiling specialized corpora where existing corpora are not deemed suitable for a given study – and develop methods for tracing meaning over time (e.g. Marchi, 2018; McEnery et al., 2019). The present "work-in-progress" report aims at contributing to this growing area of diachronic corpus studies and examines the representation of identity documents – and their use as surveillance measures (see e.g. Lyon, 2009) – focusing on the coverage of the proposed Identity Cards Scheme in the early 2000s and the more recently introduced UK Covid-19 "Vaccine Passports". There are plans to further extend this analysis to the Biometric Residence Permits required for non-EEA/UK citizens residing in the UK and the settled status application forms for EEA citizens. As social constructs, identification schemes develop out of the negotiation in discourse, for example in parliament and the media, among other contexts. National identification policies in particular are heatedly debated and subject to change, with the UK's abandoned 2006 ID card scheme being a case in point (see e.g. Whitley, 2009/2011). The introduction of vaccine passports has seen both positive and negative reactions on Twitter (Khan et al., 2022; also see McGlashan et al., 2021, on anti-vaccine discourses). The present study is motivated by the relative lack of corpus linguistic research focusing on discourses surrounding the concept of surveillance (with the exception of a handful of studies, including Branum & Charteris-Black, 2015; Elgesem & Salway, 2015, and Wiegand, 2019). It addresses the research question "What are the differences and similarities in the media coverage of different types of the UK Identity Cards Act 2006 and the NHS Vaccine passport?". The analysis extends a case study in Wiegand (2019), which examined developments in collocations across the *Times (London)'s* coverage of the Identity Cards Act 2006 from 2002–2008, by comparing the media coverage of the identity cards scheme with the coverage of the Covid-19 "Vaccine Passports" provided by the UK's National Health Service. The original case study had found new collocates on the identity card scheme emerging in line with the development of the parliamentary debate in the House of Commons and the House of Lords, for example on handling biometric data, and highlighted negative patterns in concordances in the years before the scheme was abandoned. The analysis uses data from the Times Digital Archive (Gale Cengage, 2021), the Coronavirus Corpus (Davies, 2021), and Nexis Advance (LexisNexis, 2022), for collocation, cluster, and keyword analysis. The results are expected to point to legitimization strategies related to public health concerns and emphasize data protection.

References

Branum, J., & Charteris-Black, J. (2015). The Edward Snowden affair: A corpus study of the British press. *Discourse & Communication*, *9*(2), 199–220.

Davies, M. (2021). *The Coronavirus Corpus: Design, construction, and use*. *26*(4), 583–598.

Elgesem, D., & Salway, A. (2015). *Traitor, whistleblower or hero? Moral evaluations of the Snowden-affair in the blogosphere*. Presented at the Corpus Linguistics 2015, Lancaster.

Gale Cengage. (2021). *The Times Digital Archive*. https://www.gale.com/intl/c/the-times-digital-archive

Khan, M. L., Malik, A., Ruhi, U., & Al-Busaidi, A. (2022). Conflicting attitudes: Analyzing social media data to understand the early discourse on COVID-19 passports. *Technology in Society*, *68*, 101830. https://doi.org/10.1016/j.techsoc.2021.101830

LexisNexis. (2022). *Nexis Advance*. https://advance.lexis.com/

Lyon, D. (2009). *Identifying Citizens: ID Cards as Surveillance*. Cambridge: Polity.

Marchi, A. (2018). Dividing up the data: Epistemological, methodological and practical impact of diachronic segmentation. In C. Taylor & A. Marchi (Eds.), *Corpus Approaches to Discourse: A Critical Review* (pp. 174–196). Abingdon: Routledge.

McEnery, T., Brezina, V., & Baker, H. (2019). Usage Fluctuation Analysis: A new way of analysing shifts. *International Journal of Corpus Linguistics*, *24*(4), 413–444.

McGlashan, M., Gee, M., Kehoe, A., Lawson, R., & Tkacukova, T. (2021). *TRAC:COVID Case study 2: Misinformation, authority, and trust* [Working Paper]. Birmingham: Birmingham City University. Retrieved from: http://www.open-access.bcu.ac.uk/12011/

Whitley, E. A. (2011). Perceptions of government technology, surveillance and privacy: The UK Identity Cards Scheme. In B. J. Goold & D. Neyland (Eds.), *New Directions in Surveillance and Privacy* (pp. 133–156). Abingdon: Routledge. (Original work published 2009)

Wiegand, V. (2019). *A Corpus Linguistic Approach to Meaning-Making Patterns in Surveillance Discourse* [PhD, University of Birmingham]. https://etheses.bham.ac.uk//id/eprint/9778/

Agreement with Collective Nouns in African and Caribbean Englishes

Guyanne Wilson (TU Dortmund)
guyannewilson@gmail.com

Although there has been considerable work done on agreement with collective nouns in Inner Circle Englishes (Levin 2001; 2006, Hundt 2009) and Outer Circle Englishes (Hundt 2006), African and Caribbean Englishes are often excluded completely from such discussions. Where these varieties are addressed, the treatment is rather cursory; Sand (2008) for example, looks at just eight collective nouns. This study seeks to redress this imbalance by looking at agreement with collective nouns in six varieties of English in Africa and the Caribbean: Ghanaian, Nigerian, Ugandan, Grenadian, Jamaican, and Trinidad and Tobagonian. Specifically, the main research questions guiding this study are:

- Do collective nouns in African and Caribbean Englishes exhibit singular or plural agreement?
- How does agreement with collective nouns in these varieties vary according to REGISTER and TYPE of agreement?
- What differences in agreement patterns exist for individual collective nouns?
- What differences are to be found within varieties in the same region?
- What differences in agreement with collective nouns are to be found across the two regions?

Data comprised the constituent corpora of the International Corpus of English. For the African Englishes, the corpora used were ICE Ghana, ICE Nigeria, and ICE Uganda. For all ICE corpora, both spoken and written components were used. For the Caribbean, ICE Jamaica and ICE Trinidad and Tobago were used. In addition to this, a parallel corpus of spoken and written Grenadian English was compiled.

A word list of 139 high and low frequency collective nouns was created. This word list was used to run separate concordances on each of the corpora out using AntConc. The tokens obtained were then coded for type of agreement (verbal or pronominal), register (written or spoken), and number (singular or plural). Descriptive statistics were done with the data, subsequent to which a logistic regression model was done to determine the effect of REGISTER (spoken/ written), and TYPE OF AGREEMENT (verbal/pronominal) on the use of singular agreement with collective nouns.

The African and Caribbean Englishes studied here show high rates of singular agreement, especially in written contexts, though there was some minor variation with individual lexical items, such as *staff*. Singular agreement is higher than the rates of singular agreement previously reported for British English (Levin 2001, 2006), and for both Singaporean and Philippines English (Hundt 2006), but generally slightly lower than the rates of singular agreement reported for American English. Singular verbal agreement occurs more frequently than singular pronominal agreement, particularly in spoken contexts. Indeed, in some varieties, plural pronominal agreement dominates in speech. African and Caribbean Englishes distinguish themselves from Inner Circle Englishes through the high frequency of singular pronominal agreement in written texts.

The results underscore the benefits of using ICE corpora to carry out cross-varietal studies since they allow data from geographically distant places to be compared

systematically. They also highlight the importance of including corpus-based studies African and Caribbean Englishes in the description of world Englishes.

References

Hundt, Marianne. 2006. The committee has/have decided: On concord patterns with collective nouns in inner-and outer-circle varieties of English. *Journal of English Linguistics* 34(3). 206-232.

Hundt, Marianne. 2009. Concord with collective nouns in Australian and New Zealand English. In Pam Peters, Peter Collins & Adam Smith (eds), *Comparative Studies in Australian and New Zealand English: Grammar and Beyond,* 205-222. Amsterdam: John Benjamins.

Levin, Magnus. 2001. *Agreement with collective nouns in English*. Lund: Lund Studies in English.

Levin, Magnus. 2006. Collective nouns and language change. *English Language & Linguistics* 10(2). 321-343.

Sand, Andrea. 2008. Angloversls? Concord and interrogatives in contact varieties of English. In Terttu Nevalainen, Irma Taavitsainen, Päivi Pahta & Minna Korhonen (eds.), *The dynamics of linguistic variation: Corpus evidence on English past and present*, 183-202. Amsterdam: John Benjamins.

Creole and power: A Critical Discourse Analysis of legal cross-examinations in ICE Trinidad and Tobago and ICE Jamaica

Guyanne Wilson (TU Dortmund) & Michael Westphal (University of Münster)
guyannewilson@gmail.com; michael.westphal@wwu.de

This paper examines the use of Creole and English in courtrooms in Jamaica and Trinidad. In contrast to previous research on linguistic variation in the courtroom (e.g. Rickford & King 2016; Robertson & Evans 2020), we focus not on linguistic discrimination but on the linguistic strategies with which attorneys do power (i.e. exert their authority) in this context. Traditionally, English is the language of prestige and power in the anglophone Caribbean, whereas Creoles have been described as being stigmatized and powerless (e.g. Rickford & Traugott 1985). In the paper, we address the following research questions:

- How do the attorneys use English to do power in the courtroom?
- How do they use Creole to do power?
- How they use different question structures to do power?

We discuss the findings on these questions in relation to sociolinguistic changes in the Caribbean and also highlight methodological implications of using linguistic corpora for qualitative critical analyses.

As the data for the analysis, we use legal cross-examinations (S1B 61-70) from the Trinidad and Tobago, compiled between 2007 and 2016, and Jamaica components of the International Corpus of English (ICE), compiled in the latter half of the 1990s. This very small and specific sub-corpus of 40,000 words allows us to carry out a qualitative Critical Discourse Analysis of the legal texts. In a bottom-up corpus driven analysis of the data, we identified different linguistic strategies used to do power in the courtroom. In this talk, we analyze individual interactions qualitatively to highlight how attorneys assert their authority by using English, Creole, and different question structures.

Our analysis shows that attorneys use English as their default code to display their position of authority. English dominates talk between attorneys and they predominantly use it to address witnesses, including those who respond in Creole. Hence, English is used to intimidate and silence Creole speakers. However, attorneys also code-switch to Creole when quoting witnesses verbatim and to address them directly in an antagonistic way. Attorneys use a wide range of question structures but mostly rely on questions without *do*-support or inversion to pressure witnesses to disclose information. These question types are often combined with English variant question tags (e.g. *do you*) as well as Creole tags *nah* or *eh*.

Although English remains "the language of the law" (Devonish 1986: 89), Creole has entered this formal domain. However, it is only people in position of power who can exploit Creole as strategy to do power. Creole-speaking witnesses are silenced through English.

On a methodological level, we highlight the largely untapped potential of linguistic corpora for qualitative critical analyses. Although mainly compiled for quantitative research, the ICE corpora offer the possibility for corpus-driven Critical Discourse Analyses as they are tidy and rich in background information, texts are available completely, and the ICE design covers a wide range of spoken interactive text types. Such qualitative corpus-based analyses of linguistic variation may uncover patterns often overlooked in quantitative approaches.

References

Devonish, H. (1986). *Language and liberation: Creole language politics in the Caribbean*. London: Karia Press.

Rickford, J. R. & King, S. (2016). Language and linguistics on trial: Hearing Rachel Jeantel (and other vernacular speakers) in the courtroom and beyond. *Language*, 92(4). 948-988.

Rickford, J. R., & Traugott, E. (1985). Symbol of powerlessness and degeneracy, or symbol of solidarity and truth?: Paradoxical attitudes toward Pidgins and Creoles. In S. Greenbaum (Ed.), *The English language today* (pp. 252–261). Oxford: Pergamon.

Robertson, I., & Evans, S. (2020). Systemic linguistic discrimination and disenfranchisement in the Creolophone Caribbean: The case of the St. Lucian Legal System. In E. Blake & I. Buchstaller (Eds.), *The Routledge companion to the work of John R. Rickford*. New York: Routledge.

Use of evaluative *that* in research articles: Variations across paradigms

Mei Yang (University of Helsinki)
mei.yang@helsinki.fi

The view of academic writing as evaluative and interpersonal, rather than objective and impersonal, has gained growing popularity in EAP research (Hyland & Tse, 2005a). One important interpersonal resource that academic writers can utilise to make arguments and persuade readers is evaluative *that*-clauses. The use of evaluative *that* has attracted considerable interest in research writing as the findings of this pattern may help writers effectively construct knowledge and organise discourse. However, previous research on evaluative *that* has mainly focused on variations across registers, genres, disciplines, and research article sections. What remains to be explored is how the epistemologies underlying qualitative, quantitative, and mixed methods research paradigms may affect the way academic writers signal stance via *that*-clauses.

Drawing on Hyland & Tse's (2005b) model of evaluative *that*-clauses, this study adopts a corpus-based approach to explore how the use of this structure may vary across qualitative, quantitative, and mixed methods research articles as a whole and by sections (i.e., Abstract, Introduction, Methods, Results, and Discussion). The 1,074,574-word corpus used for this study consists of 243 empirical research articles (eighty-one articles for each research paradigm) published in twenty-seven high-ranking journals during 2018-2020 in nursing. Research article sections were labelled drawing on Yang & Allison's (2004) classification of section headings (i.e., conventional headings, varied functional headings, and content headings), and a Python script was then developed to segment each article into five new files corresponding to the five sections. After segmented by section, the texts were tagged for part-of-speech (POS) with *TagAnt* (Anthony, 2015), and a Python script was then developed to extract the cases containing *that*-clauses. The raw frequencies of each category of the coded items for each file were counted and then converted into normalised frequencies per 1,000 words. The obtained data were then analysed in *R* using Kruskal-Wallis test.

The statistical test showed significant differences in all the four elements of evaluative-*that* patterns (i.e., evaluated entity, evaluative stance, evaluative source, and evaluative expression), with a large effect size noted for each element. Besides, the post-hoc pairwise comparisons showed that as a whole, qualitative RAs contained significantly more evaluative-*that* patterns than quantitative and mixed methods RAs. However, significant differences were not invariably found between qualitative and quantitative RAs, or between qualitative and mixed methods RAs across research article sections.

This study has methodological implications for future research. First, the fact that the variations in evaluative *that* across whole texts are not invariably the same as those across sections indicates that subtle local differences may be obscured when texts are examined as a whole. Thus, this study suggests that further studies divide whole texts into sections to get a nuanced understanding of paradigmatic variations. Second, the paradigmatic variations found in this study call into question how reliable the results are in an unbalanced corpus where research paradigms are not differentiated. This is important for cross-disciplinary studies because unbalanced data would make it unclear whether the variations are caused by research paradigm or by discipline. Therefore, this study recommends that the variable of research paradigms be controlled for in future cross-disciplinary studies.

References

Anthony, L. (2015). TagAnt (Version 1.2.0)[Computer Software]. *Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/*.

Hyland, K., & Tse, P. (2005a). Evaluative that constructions: Signalling stance in research abstracts. *Functions of Language*, *12*(1), 39–63. https://doi.org/10.1075/fol.12.1.03hyl

Hyland, K., & Tse, P. (2005b). Hooking the reader: A corpus study of evaluative that in abstracts. *English for Specific Purposes*, *24*(2), 123–139. https://doi.org/10.1016/j.esp.2004.02.002

Yang, R., & Allison, D. (2004). Research articles in applied linguistics: Structures from a functional perspective. *English for Specific Purposes*, *23*(3), 264–279. https://doi.org/10.1016/S0889-4906(03)00005-X.

Which factors are at play in English argument structure variation? NPs vs PPs throughout time

Eva Zehentner
eva.zehentner@es.uzh.ch

The present paper takes a broad-scale approach to cognitive factors at play in English argument structure, focussing on (changes in the) variation between verb-attached (i.e. non-subject) NP versus PP arguments. More specifically, the paper zooms into the factors impacting the choice between nominal and prepositional patterns as illustrated in (1) and (2), and discusses potential changes in such factors from Middle English to Late Modern English (1150-1914).

(1) They protested *[against] the new regulations*.
(2) They came back *[on] that day*.

Although it is generally assumed that PP-expression increased in frequency over time as part of the typological shift of English from more synthetic to more analytic (e.g. Baugh & Cable 1993), research into such variation has to day typically been restricted to narrowly-defined alternations, and a systematic, encompassing investigation into the issue is lacking so far. In this paper, I therefore retrieve all instances of verb-dependent NPs and PPs (N= approx. 406,000) from the *Penn-Helsinki Parsed Corpora of Historical English* (PPCME2, PPCEME, PPCMBE2), and annotate the arguments for position in the clause, syntactic function/ semantic role, length, and a range of morphosyntactic and semantic-pragmatic variables related to 'prominence' or markedness, such as pronominality, definiteness, or animacy (Aissen 2003; Schuhmacher & van Heusinger 2020).

These data are used as input for mixed-effects logistic regression modelling as well as conditional random forest analyses (e.g. Levshina 2015; Winter 2019), with the aim to test two opposing hypotheses about the choice between NPs and PPs: (i) a prediction grounded in accounts of 'differential argument marking', and (ii) an account emphasising the relevance of semantic role or 'function'.

Differential argument/ object marking is broadly defined as "[a]ny kind of situation where an argument of a predicate bearing the same generalized semantic argument role may be coded in different ways, depending on factors other than the argument role itself" (Witzlack-Makarevich & Seržant 2018: 3; cf. also e.g. Haspelmath 2019; Tal et al. 2020; Levshina 2021; among many others). That is, in various languages such as Spanish or Hebrew, arguments which are atypical in terms of prominence features (e.g. animate objects when objects are expected to be inanimate) are more likely to receive differential case or prepositional marking. Following this approach, I hypothesise that prominence factors may also play an important role in the choice between English NPs and PPs, and that arguments with greater prominence (e.g. animate, definite, pronominal elements) increase the probability of prepositional marking. By contrast, on the latter view, I expect NP/PP variation to be guided primarily by the arguments' basic meaning and/or syntactic properties, with more adjunctival elements (such as time or place) being more frequently expressed as PPs than more 'core' complements (e.g. recipients or themes) of verbs, which supposedly prefer NP-patterns.

I then assess these hypotheses with a particular focus on interaction with time (as well as constituent order), finding that both hold to varying degrees over time, and concluding that the history of English argument structure variation presents a case of highly complex restructuring of relations.

References

Aissen, J. 2003. Differential object marking: Iconicity vs. economy. *Nat. Lang. Linguist. ^ Theory* 21(3), 435-483.

Baugh, A. & T. Cable. 1993. *A history of the English language*. London: Routledge.

Haspelmath, M. 2019. Differential place marking and differential object marking. *STUF – Language Typology and Universals* 72(3), 313-334.

Levshina, N. 2015. *How to do linguistics with R: Data exploration and statistical analysis*. Amsterdam: Benjamins.

Levshina, N. 2021. Communicative efficiency and differential case marking: a reverse engineering approach. *Linguistics Vanguard* 7(s3), 20190087.

PPCEME = Kroch, A., B. Santorini & L. Delfs. 2004. The Penn-Helsinki Parsed Corpus of Early Modern English. Department of Linguistics, University of Pennsylvania, first edition, release 3. http://www.ling.upenn.edu/ppche/ppche-release-2016/PPCEME-RELEASE-3 .

PPCMBE2 = Kroch, A., B. Santorini & A. Diertani. 2016. The Penn Parsed Corpus of Modern British English. Department of Linguistics, University of Pennsylvania, second edition, release 1. http://www.ling.upenn.edu/ppche/ppche-release-2016/PPCMBE2-RELEASE-1 .

PPCME2 = Kroch, A., A. Taylor & B. Santorini. 2000. The Penn-Helsinki Parsed Corpus of Middle English. Department of Linguistics, University of Pennsylvania, second edition, release 4. http://www.ling.upenn.edu/ppche/ppche-release-2016/PPCME2-RELEASE-4 .

Tal, S., K. Smith, J. Culbertson, E. Grossman & I. Arnon. 2020. The impact of information structure on the emergence of differential object marking: An experimental study. *PsyArXiv*.

Van Heusinger, K. & B. Schuhmacher. 2019. Discourse prominence: Definition and application. *Journal of Pragmatics* 154, 117-127.

Winter, B. 2019. *Statistics for linguists: An introduction using R*. New York, NY: Routledge.

Witzlack-Makarevich, A. & I. Seržant. 2018. Differential argument marking: Patterns of variation. In Seržant, I. & A. Witzlack-Makarevich (eds.), *Diachrony of differential argument marking*, 1-40. Berlin: Language Science Press.

Posters

Connections between genres and error patterns in a Swedish upper-secondary EAL learner corpus

Daniel Ihrmark (Linnaeus University)
daniel.o.sundberg@lnu.se

Genre in second language (L2) writing can be seen as frames from which students can learn to draw conclusions about how to communicate within a certain domain (Hyland, 2004, p. 3). A central component is the idea of a discourse community within which the students expect to find their intended reader, and the features of the genre that would fit within that community (Hyland, 2019, p. 24), thus aiding in their choice of linguistic features (Riley and Reedy, 2000). The influence of the genre frame can be seen to differ between second language learners and foreign language learners, where the latter rely more heavily on the conventions provided by the genre (Yasuda, 2011). This makes English in the Swedish context interesting due to the transculturation of the language, placing it in an interesting position vis-à-vis foreign and second language (Hult, 2012).

Taking this approach to genres in writing as an influence on students' language choices, the poster proposed here intends to answer the question of whether or not the choices brought on by genre result in different sets of errors in student writing.

By applying the Java LanguageTool (JLT) to a corpus of Swedish upper-secondary learner texts, the proposed poster intends to look for connections between the error categories provided by JLT in relation to the writing genre of the collected texts. JLT is based on the work of Naber (2003), which is an open-source grammar and style checker based on POS-tagging and chunking materials for comparison with a pre-defined set of error patterns. This paper will make use of the Python implementation by Morris (2020).

The texts for the learner corpus were collected from an anonymous Swedish forum where students upload their written production for community feedback. Uploads are categorized according to school subject and grade level (primary, secondary or upper-secondary). Collection was done using a BeautifulSoup4 web crawler, and the collected texts were tagged for genre manually. In total, the learner corpus contains 470 texts distributed over 9 text genres. The most common genres, in order of frequency, were reports, informative essays, short fiction, biographies and argumentative essays. Letters, lyrics, poetry and speeches were also represented, but at much lower frequencies.

Due to the distribution of the genres, the expected results for the proposed poster are the distributions of error patterns amongst the frequent text genres in Swedish EAL learners' written production. The thesis for the poster is that the conventions of the genres would lead students to produce texts which exhibit error patterns resulting from the language choices made based on said conventions. If this is the case, information about which error patterns are frequent in which genres will allow teachers to proactively plan their lessons to inform about relevant linguistic features before a writing task involving a specific genre.

Reference

Hult, F.M. (2012) 'English as a Transcultural Language in Swedish Policy and Practice', TESOL Quarterly, 46(2), pp. 230–257. doi:10.1002/tesq.19.

Hyland, K. (2004) Genre and second language writing. Ann Arbor: University of Michigan Press (Michigan series on teaching multilingual writers).

Hyland, K. (2019) Second language writing. Second edition. Cambridge, United Kingdom ; New York, NY: Cambridge University Press.

Morris, J. (2020) language-tool-python 2.7.0. Python Software Foundation. Available at: https://pypi.org/project/language-tool-python (Accessed: 28 January 2022).

Naber, D. A. (2003) “Rule-based Style and Grammar Checker”, Diplomarbeit, Universität Bielefeld, Bielefeld. Available at: http://www.danielnaber.de/languagetool/download/style and grammar checker.pdf

Riley, J. and Reedy, D. (2000) Developing writing for different purposes: Teaching about genre in the early years. Sage.

Yasuda, S. (2011) ‘Genre-based tasks in foreign language writing: Developing writers’ genre awareness, linguistic knowledge, and writing competence’, Journal of Second Language Writing, 20(2), pp. 111–133. doi:10.1016/j.jslw.2011.03.001.

The design of a Corpus of late Modern English Texts on Physics

Luis Puente-Castelo (Universidade da Coruña), pcastelo.luis@gmail.com
Leida Maria Monaco (University of Oviedo), mariamonaco86@gmail.com
Isabel Moskowich (Universidade da Coruña), imoskowich@udc.es
Begoña Crespo (Universidade da Coruña),begona.crespo.garcia@udc.es

This poster aims at presenting CETePh, Corpus of English Texts on Physics, a new component of the Coruña Corpus of English Scientific Writing (CC). CETePh is being compiled with the view of describing the language used in the discipline of Physics during a period in which the different scientific registers of English were being developed.

One of the distinctive characteristics of the CC has always been the gathering of late Modern English scientific texts as several twin corpora sharing identical principles (Monaco and Puente-Castelo, 2019; Moskowich 2019; Crespo and Moskowich, 2020). The first step of the compilation involved the delimitation of the time-span to be covered and the size of the samples gathered. There are many different proposals for the periodisation of late Modern English, but we have decided to accept 1700 and 1900 as time limits for all disciplines, considering the situation of science as well as some historical events in this period. As for sample size, and following the general compilation principles applied to all the subcorpora in the Coruña Corpus (Crespo and Moskowich, 2009; Crespo and Moskowich, 2020), we opted for the collection of 10,000-word extracts. At the moment of making this decision, specialised corpora were just a few and not much information about them was available, except for Biber's (1993) claim that variation in specialised registers could be detected and studied in 1,000-word samples. However, both our close reading of the material and the difficulty to collect large amounts of extracts convinced us to select 10,000 words as the definite size for the samples collected. This decision, together with our determination to use XML-TEI both for text and metadata files, are now supported by other corpus compilers (VARIENG, 2016) that have partly adopted similar decisions.

The Coruña Corpus has been organised into different sub-corpora, one per discipline, with the intention of representing late Modern specialised texts. Our delimitation of disciplines was based on the criteria of the historical moment in which texts were published, that is, we adopted an inclusive perspective. All in all, as a starting point for discipline categorisation, we have used the divisions proposed by UNESCO in 1988. This way, several subcorpora have already been published: The Corpus of English Texts on Astronomy (CETA, 2012) and the Corpus of English Philosophy Texts (CEPhiT, 2016), the Corpus of History English Texts (CHET, 2019), the Corpus of English Life Sciences Texts (CELiST, 2020) and the Corpus of English Chemistry Texts (CECheT, 2022). Others, however, are found in different stages of the compilation process. This is the case of the Corpus of English Texts on Languages (CETeL) and, more recently, the Corpus of English Texts on Physics (CETePh), the object of study of this poster. Thus, one of the aspects that we will deal with here has to do with the description of the discipline itself: its evolution and some contextual peculiarities of physics in the eighteenth and nineteenth centuries.

In view of all this, CETePh – as well as every other subcorpus of the CC – is built in a way that each text sample is accompanied by a metadata file. Such files provide information both about the author and the text itself. These metadata files, in combination with the Coruña Corpus Tool, can be used to narrow searches according to extralinguistic parameters (sex, age or geographical provenance of the author as well as date of publication, genre of the sample, etc.). This poster will present the contents of these

metadata files, focusing on each aspect that may be considered relevant for the description of the corpus, in that they all embody the idiosyncrasy of CETePh.

References

Biber, Douglas. (1993). Representativeness in corpus design. Literary and Linguistic Computing, 8, 243-257.

Crespo, Begoña and Moskowich, Isabel. (2009). "CETA in the context of the Coruña Corpus". Literary and Linguistic Computing, 25: 2, 153-164.

Crespo, Begoña. & Moskowich, Isabel. (2020). "Astronomy, philosophy, life sciences and history texts: setting the scene for the study of modern scientific writing". English Studies, 101: 6, 665-684.

Lareo, Inés; Monaco, Leida Maria; Esteve-Ramos, María-José; Moskowich, Isabel (comps.) (2020). *Corpus of English Life Sciences Texts*. A Coruña: Universidade da Coruña. https://doi.org/10.17979/spudc.9788497497848

LMEMT. (n.d.). Retrieved 24 June, 2016, from VARIENG research Unit for variation, contacts and change in English. http://www.helsinki.fi/varieng/CoRD/corpora/CEEM/LMEMTindex.html

Monaco, Leida Maria & Puente-Castelo, Luis. (2019). "'A matter both of curioſity and uſefulneſs': Compiling the Corpus of English Texts on Language". Research in Corpus Linguistics, 7, 47-68.

Moskowich, Isabel. (2019). "An introduction to CHET, the Corpus of History English Texts". In Moskowich, Isabel; Begoña Crespo, Luis Puente-Castelo & Leida Maria Monaco (eds.) *Writing history in Late Modern English: Explorations of the Coruña Corpus*, 42-56.

Moskowich, Isabel; Lareo, Inés; Lojo Sandino, Paula & Sánchez-Barreiro, Estefanía (comps.) 2019. *Corpus of History English Texts*. A Coruña: Universidade da Coruña. https://doi.org/10.17979/spudc.9788497497091

Moskowich, Isabel; Puente-Castelo, Luis & Monaco, Leida Maria (comps.) 2022. *Corpus of English Chemistry Texts.* A Coruña: Universidade da Coruña. https://doi.org/10.17979/spudc.9788497498388

Moskowich, Isabel & Crespo, Begoña (2012). Corpus of English Texts on Astronomy (CETA). [CD-Rom] Included in *Astronomy 'playne and simple'. The writing of science between 1700 and 1900*. Amsterdam: John Benjamins.

Moskowich, Isabel & Crespo, Begoña (2016). "Classifying communicative formats in CHET, CECHeT and others", EPiC Series in Language and Linguistics, 1, 308-320.

Moskowich, Isabel; Camiña-Rioboo, Gonzalo; Lareo, Inés & Crespo, Begoña (2016). Corpus of English Philosophy Texts (CEPhiT). [CD-Rom] Included in *'The Conditioned and the Unconditioned'. Late Modern English texts on philosophy*. Amsterdam: John Benjamins.

Puente-Castelo, Luis & Monaco, Leida Maria (2016). "'it is proper subserviently, to inquire into the nature of experimental chemistry': Difficulties to harmonize disciplinary particularities and compilation criteria during the selection of samples for CECheT". EPiC Series in Language and Linguistics, 1, 351-360.

UNESCO (1988). *Proposed international standard nomenclature for fields of science and technology*. Paris: UNESCO.

The impact of *Star Wars* on the English language: A study of *Star Wars*-derived words and constructions in present-day English corpora

Christina Sanchez-Stockhammer (TU Chemnitz)
christina.sanchez@phil.tu-chemnitz.de

Since George Lucas' film *A New Hope* was first screened in 1977, the *Star Wars* series has become much more than a simple world-wide box-office success: this pop-culture phenomenon now incorporates films, computer games, books and merchandise, and the most ardent fans even understand *Star Wars* as a quasi-religious philosophy (cf. Davidsen 2016).

Surprisingly, however, linguistic research related to *Star Wars* is extremely scarce and only a very recent phenomenon, with a focus on language use in the films themselves (cf. Yuliyana & Bram's 2019 syntactic analysis of Master Yoda's unusual word order, Allen's 2019 onomastic analysis of the characters' names in the spin-off *Star Wars Rebels* or Kleiven's 2021 study of attitudinal language use). In view of the extremely large fan base, which has already begun to accommodate the second generation of viewers, it is particularly surprising that there is as yet no general investigation of the impact of *Star Wars* on the English language. The present contribution fills this gap by using corpus-linguistic methods to investigate the extent to which characteristic words and constructions from the *Star Wars* universe can be considered established in the English language.

The starting point consisted in three *Star Wars*-related words that the *Oxford English Dictionary* (www.oed.com) included as headwords in its October 2019 update (https://public.oed.com/updates/), and which received media coverage, namely *Jedi*, *Padawan* and *lightsabre* (with spelling variants). Following a full-text search for *Star Wars* in the OED database, this list was complemented by *Yoda* (full entry since 2016) and *the dark side* (full entry since December 2021) in the characteristic construction *to the dark side*, which generates less noise in the corpus searches. Our assumption was that the establishment of a lexical item from popular culture in the English language should find its reflection not only in its general frequency of use, but also in its use without direct reference to the original films and books or derived merchandise.

In order to analyse the status of the language items in question, we carried out frequency and collocation analyses using mutual information in COCA, COHA, BNC and BNC Spoken 2014 and coded *Star Wars*-specific vs. general-language use for subsets of the data.

Our results suggest that a relevant proportion of the corpus hits for the investigated *Star Wars*-derived vocabulary and constructions do not directly refer to the *Star Wars* universe (as would e.g. be the case in mentions of the film title *Return of the* ***Jedi***). Instead, they involve more indirect references to their original source (as in the COCA example: *And before you could make a bad* ***Jedi*** *mind trick joke, we were back on the road to Lesotho.*) or correspond to completely new usage contexts that only remotely allude to their original use (as in the BNC example *Other imbibers have gone over* ***to the dark side*** *of beer, rejecting the pasteurised lager produced by the breweries.*). The analyses thus suggest the ongoing integration of *Star Wars*-derived words and constructions into the English language.

References

Allen, Spencer L. 2019. A rebel by any other name: The onomastics of Disney's Star Wars Rebels. *The Journal of Religion and Popular Culture* 31:1. 72-86.

Davidsen, Markus Altena. 2016. From Star Wars to Jediism: The emergence of fiction-based religion. In E. van den Hemel & A. Szafraniec (eds.), *The future of the religious past*, 376-389. Fordham University Press.

Kleiven, Ragnhild Fimreite. 2021. *May the accent be with you: An attitudinal study of language use in the Star Wars trilogies*. University of Bergen: Master thesis. https://hdl.handle.net/11250/2760489.

Yuliyana, Yuliyana & Barli Bram. 2019. Uncommon word order of Yoda in Star Wars movie series: A syntactic analysis. *NOBEL: Journal of Literature and Language Teaching* 10:2. 103-116.

Software Demos

Software demonstration: meaning-based querying of historical corpora with MacBERTh

Lauren Fonteyn & Enrique Manjavacasl.fonteyn@hum.leidenuniv.nl;
e.m.a.manjavacas.arevalo@hum.leidenuniv.nl

Because of great efforts in corpus linguistics, a large number of historical texts have been digitized (and sometimes even syntactically parsed and part-of-speech-tagged), which has enabled the automatic retrieval of words/phrases/sentence structures by means of formal queries. The next step in corpus querying, then, would be to move from formal querying to semantic querying, but this has proven a difficult challenge in the past. However, in recent years, it became evident that recent advanced in Distributional Semantic models, from type-embeddings derived from algorithms like word2vec (Mikolov et al, 2013) to contextualized token-embeddings such as BERT (Devlin et al. 2019), can help capture the denotations and connotations of linguistic items. Contextualized embeddings in particular perform excellently in (semi-)automatic sense disambiguation and exemplar-based retrieval (e.g. Fonteyn 2020 for an application to a linguistic case study).

This software demonstration will focus on MacBERTh, a BERT-based model pre-trained on Early Modern and Late Modern English (3.9B (tokenized) words, time span: 1450-1950; Manjavacas & Fonteyn 2021, 2022). We will demonstrate how MacBERTh may help researchers (i) access and (ii) analyse the semantic information encoded in linguistic corpus data in a (semi-)automatic way. Because the contextualized embeddings MacBERTh produces can be used to measure semantic 'closeness' between word/phrase/sentence tokens, they can be manipulated to enable semantic searches within textual material:

target	>> Mr Reynolds called on me about the drawing of the Laird `s jock (1825) << (OED sense: 'regarding')	
rank	**date**	**Nearest neighbours in corpus**
1	1818	*I have spoken to Haydon about the drawing*
2	1735	*I called at Hammersmith yesterday about Lady Ailesbury `s tubs*
…	…	
target	>> The average was about fourteen (1885) << (OED sense: 'approximately'	
rank	**date**	**Nearest neighbours in corpus**
1	1832	The average price appears to be about 75 s
2	1883	The average age of the children was about 3 years
…	…	
target	>> There is frequently much selfishness about remorse . (1852) << (OED sense: 'connected with', figurative)	
rank	**date**	**Nearest neighbours in corpus**
1	1903	*Why should there be this queer flavour of absurdity and pretentiousness about the thing ?*
2	1910	*There is this element of a fine fruitfulness about the male enjoyments*
…	…	

Table 1 - Example of sense retrieval task for 'about' with MacBERTh. Approximately 40,000 examples of 'about' (and all spelling variants listed in the Oxford English Dictionary) were retrieved from CLMET3.1. For each target example, a list of examples is generated ranked according to similarity (metric: cosine).

By creating embeddings with MacBERTh, researchers will also be able to map out the distances between senses of linguistic items (in different texts and different time stages) as formalized in their representational distances. In this demonstration, we walk participants through two case studies (i.e. lexical and grammatical) that show how MacBERTh can support historical text

analysis (from data collection to pre-processing to (error) analysis). During the demonstration, participants are guided through the open source code notebooks (in jupyter) and instructional material (which will be made available through the project website https://macberth.netlify.app/). Participants are also encouraged to discuss possible research questions that MacBERTh (or its Dutch sibling GysBERT) may help address.

References

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*, 4171–86. Minneapolis, Minnesota.

Fonteyn, Lauren. 2020. *Let's get into it: Using contextualized embeddings as retrieval tools*. In: Timothy Colleman, Frank Brisard, Astrid De Wit, Renata Enghels, Nikos Koutsoukos, Tanja Mortelmans and María Sol Sansiñena (eds.), The Wealth and Breadth of Construction-Based Research [Belgian Journal of Linguistics 34], 66–78.

Manjavacas Arévalo, Enrique & Lauren Fonteyn. 2021. MacBERTh: Development and evaluation of a historically pre-trained language model for English (1450-1950). In *Proceedings of the workshop on natural language processing for digital humanities (nlp4dh)*, 23–36. Association for Computational Linguistics.

Manjavacas Arévalo, Enrique & Lauren Fonteyn. 2022. Adapting vs Pre-training Language Models for Historical Languages. *Journal of Data Mining and Digital Humanities*.

Mikolov, Tomás, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. 'Efficient Estimation of Word

Representations in Vector Space'. In Yoshua Bengio and Yann LeCun (eds.), *1st International Conference on Learning Representations, ICLR 2013*, Scottsdale, Arizona, USA, May 2-4, 2013. http://arxiv.org/abs/1301.3781.