The Twelfth International Corpus Linguistics Conference

2023

**Book of Abstracts**

****

**Monday 3rd - Thursday 6th July**

Contents

[Titles by programme 3](#_Toc137468797)

[Monday 3rd July 3](#_Toc137468798)

[Tuesday 4th July 2023 7](#_Toc137468799)

[Wednesday 5th July 12](#_Toc137468800)

[Thursday 6th July 17](#_Toc137468801)

[Poster titles (by room order) 20](#_Toc137468802)

[Abstracts: Panels 23](#_Toc137468803)

[Abstracts: Papers 30](#_Toc137468804)

[Abstracts: Posters 154](#_Toc137468805)

# Titles by programme

## Monday 3rd July

9.00 Welcome

**9.30 Plenary: Marc Alexander, University of Glasgow** GF LT1

10.30 Coffee

|  |  |  |  |
| --- | --- | --- | --- |
| *Time* | *Unique ID* | *Paper title and authors* | *Room* |
| 11.00-12.30 | [5085] | **PANEL:** What are Fundamental Principles of Corpus Linguistics? *Vaclav Brezina and Tony McEnery* | GF LT1 |
| 11.00 | [5446] | Examining corpus prototypicality and keyness beyond the lexical level: Experiments with ProtAnt *Nicholas Smith, Laurence Anthony, Sebastian Hoffmann and Paul Rayson* | GF LT2 |
| [9209] | The language of children's online disclosures of abuse *Mark McGlashan* | GF LT3 |
| [9245] | How reliable is our reference? The effect of the reference corpus on measuring lexical sophistication in L2 English speech *Raffaella Bottini and Elen Le Foll* | GF LT4 |
| [5255] | Gender on Dimensions of Linguistic Variation. Making Sense of Non-significance *Marianna Gracheva* | GF LT6 |
| [2142] | Bringing corpus use into classroom: The case of a Secondary School Teacher in an EFL context *Ece Genç-Yöntem and Evrim Eveyik-Aydın* | CC A15 |
| 11.30 | [9513] | A multidimensional text typology of American film *Marcia Veirano Pinto and Tony Berber Sardinha* | GF LT2 |
| [2610] | An intersectional corpus-assisted discourse study: The discursive construction of sexual misconduct in the Canadian Armed Forces *Rachelle Vessey* | GF LT3 |
| [2943] | Exploring the predictive power of multiple lexical diversity measures for L2 Spanish writing proficiency *Earl Brown, Brett Hashimoto and Alan Brown* | GF LT4 |
| [9944] | A corpus-driven study of proxies for register levelling in Present-Day British English *Laura Abalo-Dieste* | GF LT6 |
| [9089] | Effects of Data-Driven Learning on Memorizing Verb-Noun Collocations *Yoshiho Satake* | CC A15 |
| 12.00 | [972] | Dealing with Duplicates in Newspaper Corpora *Sharon Glaas* | GF LT2 |
| [6984] | Abusive name-calling: Representations around the term genocide during the Covid pandemic *Yara Maria de Toledo Dias Romeiro, Paula Tavares Pinto, Marcos Oliveira and Natalia Luri Arimori Ribeiro* | GF LT3 |
| [8695] | The effect of English extramural activities on L2 students’ lexical diversity and grammatical complexity *Henrik Kaatari, Tove Larsson, Ying Wang, Seda Acikara Eickhoff and Pia Sundqvist* | GF LT4 |
| [9642] | Linguistic and Situational Granularity within Registers: A Cross-register Perspective *Marianna Gracheva and Jesse Egbert* | GF LT6 |
| [6225] | Data Driven Learning of Sino-Japanese and Sino-Korean vocabulary *Sara Librenjak* | CC A15 |

12.30 Lunch

|  |  |  |  |
| --- | --- | --- | --- |
| 13.30-15.00 | [1200] | **PANEL:** Dead ends, detours, and disappointments: Looking beyond 'failed' corpus research *Gerlinde Mautner, Mathew Gillings, Paul Baker, Charlotte Taylor, Clyde Ancarno and Susan Hunston* | GF LT1 |
| 13.30 | [5469] | Evaluating collocation in spoken dialogic corpora *Robbie Love, Isobelle Clarke and Mark McGlashan* | GF LT2 |
| [362] | English binominal noun phrases as seen through their Czech counterparts *Gabriela Brůhová and Kateřina Vašků* | GF LT3 |
| [7148] | Coding Discourse marker functions in corpora of spoken English *Nuha Alharbi* | GF LT4 |
| [1117] | Building a corpus of academic writing in EMI contexts: Challenges of collecting data from different educational settings *Dana Gablasova, Luke Harding, Raffaella Bottini, Sally Ren and Vaclav Brezina* | GF LT6 |
| [9530] | From specialized corpus to the EAP classroom: Integrating authentic data into materials design *Ana Luiza Pires de Freitas, Ana Eliza Pereira Bocorny, Rozane Rodrigues Rebechi and Simone Sarmento* | CC A15 |
| 14.00 | [3763] | Two approaches to collocation networks: GraphColl and LLN *Hanna Schmück and Vaclav Brezina* | GF LT2 |
| [2182] | On register-specificity of personification in Hungarian – a corpus-driven study *Gábor Simon* | GF LT3 |
| [4548] | Language Learners’ Use of Conversational Persian: Insights from a Learner Corpus *Sepideh Daghbandan* | GF LT4 |
| [8310] | Using a Representativeness Argument for Corpus Evaluation *Jenny Kemp* | GF LT6 |
| [2632] | Discipline-specific lexical frames: Linguistic data extracted from corpora to inform EAP teaching across and within disciplines *Vivian Lameira, Ana Eliza Bocorny, Rozane Rebechi, Deise Prina Dutra and Tony Berber Sardinha* | CC A15 |
| 14.30 | [9196] | Towards operationalization of the concept of discursive salience as a factor in the development of semantic prosody: the case of abortion in English Web 2020 *Beth Malory* | GF LT2 |
| [7911] | Light verbs in L2 Czech *Andrea Hudousková* | GF LT3 |
| [7356] | Developing a youth talk corpus for Turkish: The Corpus of Turkish Youth Language (CoTY) and the use of vocatives as interactional markers *Esranur Efeoğlu-Özcan and Hale Işık-Güler* | GF LT4 |
| [5153] | Authorial presence across Medicine RA sections: A corpus-based study of self-mentions *Heng Gong and Michael Barlow* | GF LT6 |
| [8660] | Perceptions of Using Corpora to Enhance L2 Interactional Competence: A Study of Chinese L1 Speakers *Sarah Warfield* | CC A15 |

15.00 Afternoon tea

|  |  |  |  |
| --- | --- | --- | --- |
| 15.30-17.00 | [5327] | **PANEL:** Corpus linguistic approaches to school-aged children’s writing *Philip Durrant, Hildegunn Dirdal, Yaling Hsiao and Victorina Gonzalez-Diaz* | GF LT1 |
| 15.30 | [7236] | Progress, innovations, and challenges to date in the compilation of LANA, a corpus of American English *Elizabeth Hanks, Tony McEnery, Vaclav Brezina, Jesse Egbert, Tove Larsson, Doug Biber, Randi Reppen, Paul Baker, Gavin Brookes, Raffaella Bottini and Isobelle Clarke* | GF LT2 |
| [1624] | Historical and Contemporary Perspectives of Disinformation *William Dance* | GF LT3 |
| [2095] | “Get the woke intern away from the twitter account stat!!!” Exploring reactions to Ryanair’s digital communication *Ursula Lutzky and Andrew Kehoe* | GF LT4 |
| [3294] | The semantic prosodies of different senses of a word *Mathias Russnes* | CC A15 |
| 16.00 | [963] | Construction and Analysis of a Map-Based Corpus for Tracking Linguistic Variation and Demographic Characteristic Identification *Theodore Manning, Eugenia Lukin, Ross Klein and Patrick Juola* | GF LT2 |
| [7223] | Linguistic characteristics of pseudo-science: A corpus-based study *Deise Prina Dutra, Ana Bocorny, Maria Claudia Delfino, Paula Tavares Pinto, Simone Sarmento and Tony Berber Sardinha* | GF LT3 |
| [1649] | Shame on you! A corpus-based analysis of green marketing discourse in mission statements of certified sustainable fashion firms *Yingnian Tao and Mark Ryan* | GF LT4 |
| [2259] | Building a taxonomy of pattern-based constructions *Susan Hunston* | CC A15 |
| 16.30 | [4928] | Balance and Representativeness in the Compilation of the National Corpus of Irish (CNG) *Mícheál J. Ó Meachair, Úna Bhreathnach, Brian Ó Raghallaigh and Gearóid Ó Cleircín* | GF LT2 |
| [7455] | The Discourses of Climate Change Denialism across Conspiracy and Pseudoscience Websites *Isobelle Clarke* | GF LT3 |
| [9004] | Corpus approaches to transparency in corporate discourse - An analysis of texts from websites of companies in the car transportation sector *Federico Zaupa* | GF LT4 |
| [9137] | Application of collostructional analysis to describe linguistic remix *Emese K. Molnár* | CC A15 |
| 17.00 | [128] | Web corpora for under-resourced languages *Vit Suchomel, Miloš Jakubíček and Ondřej Matuška* | GF LT2 |
| [8643] | Distinguishing False News Through Lexico-Grammatical Features: A Multi-dimensional Analysis *Bashayer Baissa, Matteo Fuoli and Jack Grieve* | GF LT3 |

## Tuesday 4th July 2023

**9.00 Plenary: Anne O’Keeffe, Mary Immaculate College** GF LT1

|  |  |  |  |
| --- | --- | --- | --- |
| 10.00 | [2781] | National collections as corpora: Insights from the Our Heritage, Our Stories project *Ewan Hannaford and Marc Alexander* | GF LT3 |
| [4188] | Variation in Response Structure in a Corpus of Online Chat *Jack Grieve, Sara Bartl, Solly Elstein, Kira Ogburn and Dana Roemling* | GF LT4 |
| [1537] | Migrating across times and cultures. Metaphorical images of migration in U.S.A and Italian newspaper discourse between 20th and 21st century. *Dario Del Fante* | GF LT6 |
| [7112] | Text-level measures of lexical dispersion: Robustness analysis *Lukas Sönning* | CC A15 |

10.30 Coffee

|  |  |  |  |
| --- | --- | --- | --- |
| 11.00-12.30 | [6712] | **PANEL:** Discourse units and sub-register variation in British English conversation *Jesse Egbert, Douglas Biber and Tony McEnery* | GF LT1 |
| 11.00 | [567] | Creating and analysing a multimodal corpus of obesity news texts with Google Cloud *Vision*’s automatic image tagger *Luke Collins and Paul Baker* | GF LT2 |
| [8730] | A corpus of Victorian anti-vaccination discourse *Alice Deignan, Tara Coltman-Patel, William Dance, Zsófia Demjén, Derek Gatherer, Claire Hardaker, Chris Sanderson and Elena Semino* | GF LT3 |
| [7315] | Corpus-based Bilingual Terminology Extraction using One-Click Terms *Miloš Jakubíček, Ondrej Matuska and Marek Blahuš* | GF LT4 |
| [2735] | Upon my going into a coffee-house yesterday, and lending an ear to the next table…: A corpus-based exploration of Coffeehouse dialogues and their discursive practices in the period 1660-1712 *Ersilia Incelli* | GF LT6 |
| [387] | Textual Functions and Aspects of Underspecification of Discourse Markers in Translation: A Parallel Corpus-based Investigation of English and Persian *Zahra Ghane and Mehrdaad Vasheghani Farahani* | CC A15 |
| 11.30 | [9785] | A new method for coding multimodal corpora: The discursive construction of free-range animals on Twitter *Yuze Sha and Beth Malory* | GF LT2 |
| [4633] | Discourses emerging from anti-vaccine movements in Brazil on Twitter, during the COVID-19 pandemic *Mirella Whiteman, Arianne Brogini, Stephan Hughes and Paula Tavares Pinto* | GF LT3 |
| [8055] | Exploring the Transfer of Topic-Comment Structures in L2 English: A Comparison of Asian L1 Speakers *Yasuhiro Fujiwara, Takanori Iwao, Hajime Ito and Kiyoshi Naka* | GF LT4 |
| [618] | Building blocks of fairy tale openings and closings in the TransGrimm Corpus *Christina Sanchez-Stockhammer and Asya Yurchenko* | GF LT6 |
| [1317] | Variation in learner writing and a comparison with textbook models *Marilisa Shimazumi and Tony Berber Sardinha* | CC A15 |
| 12.00 | [1139] | More than words: Multimodal multidimensional analysis of popular music in English language *Maria Claudia Delfino* | GF LT2 |
| [7324] | ‘Roe was egregiously wrong from the start’: A Contrastive Analysis of the Attitudinal Evaluation of Abortion Case Law by SCOTUS Opinion Writers. *Jamie McKeown and Mary Ye* | GF LT3 |
| [8648] | Code-switching in Tunisian Arabic: A multifactorial Random Forest analysis *Chadi Ben Youssef and Stefan Th. Gries* | GF LT4 |
| [2520] | Us investigate fragments today? Why why- and Mad-Magazine fragments in contemporary English? *Yolanda Fernández-Pena and Javier Pérez-Guerra* | GF LT6 |

12.30 Lunch

|  |  |  |  |
| --- | --- | --- | --- |
| 13.30 | [3940] | Women in Early Modern Manuscript News: Harnessing a Lost Corpus with Bootstrapping and Bayesian Statistics *Gabor Mihaly Toth* | GF LT1 |
| [1549] | Discourses of police and policing in newspaper coverage of UK climate change protests, 2019-2022. *Charlotte-Rose Kennedy and David Wright* | GF LT2 |
| [3593] | Things That Make Us Go Ooh… (Why We Keep Rolling Up to the Circus) *Katharine Kavanagh* | GF LT3 |
| [9876] | Can they learn from their own mistakes? An investigation into the application of learner corpora in an English academic writing course for L2 undergraduate students *Meilin Chen* | GF LT4 |
| [6210] | Linguistic Variation within Registers – Communicative Freedom Meets Situational Constraints *Marianna Gracheva* | GF LT6 |
| [5584] | The emergence of 'publicly': Using a historical corpus to investigate an orthographic anomaly *John Williams* | CC A15 |
| 14.00 | [7162] | Down-sampling from hierarchically structured corpus data: The case of third-person verb inflection in Early Modern English *Lukas Sönning and Ole Schützler* | GF LT1 |
| [8190] | Deconstructing Discourses – A network-based approach to analysing Lone Wolf Terrorism in British Newspapers *Hanna Schmück* | GF LT2 |
| [4986] | The language of fashion from a multi-dimensional perspective *Katherine Oliva Ortolani* | GF LT3 |
| [7374] | Cross-validation across multivariate analytical systems: A statistical method for the CEFR-level estimation and automated assessment of the syntactic/lexical complexity of English-L2 learner texts *Andrea Agnes Remenyi and Karlygash Adamova* | GF LT4 |
| [3820] | Why do function word frequencies vary across individuals? Evidence in favour of The Statistical Approximation Hypothesis *Andrea Nini* | GF LT6 |
| [5366] | Can a corpus analysis of “such… as” constructions in 18th century American English facilitate interpretation of the U.S. Constitution’s Appointments Clause? *Haoshan Ren, Clark D. Cunningham and Ute Römer* | CC A15 |
| 14.30 | [6940] | Spelling regularisation for Early Modern English texts: advances in supervised and semi-supervised postediting *Andrew Hardie* | GF LT1 |
| [322] | Discursive construction of key issues in Australian citizenship legislative enquiry from 2007 to 2017: A corpus-assisted discourse analysis *Remart Dumlao* | GF LT2 |
| [2928] | The verbal language of visual arts: A multi-dimensional analysis of the discourse on Sally Mann's photography *Yara Maria de Toledo Dias Romeiro* | GF LT3 |
| [1498] | An application of Learner Corpus Αnalysis in grammatical comprehension *Katerina Florou* | GF LT4 |
| [9510] | Exploring individual variation in constructional schematicity using random effects *Svetlana Vetchinnikova* | GF LT6 |
| [154] | Network of Grammar Lexemes in 16th- and 17th-Century English Grammar Writing *Beatrix Busse, Nina Dumrukcic, Sophie Du Bois and Ingo Kleiber* | CC A15 |

15.00 Afternoon tea

**15.30 Poster session** GF LT5

|  |  |  |  |
| --- | --- | --- | --- |
| 16.00 | [9025] | Overhauling keyness analyses with three dimensions & word2vec: Contrasting three Asian varieties of English *Stefan Th. Gries* | GF LT1 |
| [7381] | From Argentina to Zimbabwe: Exploring the global appeal of the International Baccalaureate 1977-2019 *Saira Fitzgerald* | GF LT2 |
| [7346] | Indicating engagement in online workplace meetings: The role of backchannelling head nods *Anne O'Keeffe, Dawn Knight, Christopher Fitzgerald, Justin McNamara, Geraldine Mark, Svenja Adolphs, Sandrine Peraldi, Tania Fahey Palma, Fiona Farr and Benjamin Cowan* | GF LT3 |
| [6055] | Teaching spoken features with the American Movie Corpus: A task-based approach *Valentina Morgana* | GF LT4 |
| [8067] | Self-identification in GenZ environmental associations: a corpus assisted analysis of first-person plural pro-forms *Mariasophia Falcone* | CC A15 |
| 16.30 | [1601] | Computer-assisted concordance reading *Stephanie Evert, Natalie Finlayson, Michaela Mahlberg and Alexander Piperski* | GF LT1 |
| [3163] | How expert are universities? Investigating research mobilities in primary literacy education in England *Anna Lindroos Čermáková and Julia Gillen* | GF LT2 |
| [1234] | A corpus pragmatic analysis of politeness in spoken English business communication *Isolde van Dorst* | GF LT3 |
| [7545] | Corpus linguistics for audiovisual translation revisited: Netflix as a multimedia corpus *Judyta Mężyk* | GF LT4 |
| [4374] | Gayborhoods, Queeristan, and bi-nosaurs – the conceptual blending of queer terms in The TV Corpus *Charlotte Rosenow* | CC A15 |
| 17.00 | [5213] | The hierarchy of web pages: Accounting for contents’ accessibility in keywords analysis *Matteo Di Cristofaro* | GF LT1 |
| [8263] | Professional nouns in a positive light: How are internationally standardized groups reflected in news media? *Irene Elmerot* | GF LT2 |
| [5383] | Business discourse in the Irish technology sector: A corpus-based investigation of spoken language in International Virtual Teams (IVTs) *Gail Flanagan* | GF LT3 |
| [6161] | Movie transcriptions and subtitles: “Two of a kind” for spoken lexicogrammar *Marcia Veirano Pinto and Pierfranca Forchini* | GF LT4 |
| [4735] | “Be one with your external iTunes library, young Padawan.” Star Wars-derived words and constructions in English corpora and dictionaries *Christina Sanchez-Stockhammer* | CC A15 |

## Wednesday 5th July

**9.00 Plenary: Sylvia Jaworska, University of Reading** GF LT1

|  |  |  |  |
| --- | --- | --- | --- |
| 10.00 | [8062] | Parliamentary language profiles. How do the politicians speak? *Martina Berrocal and Václav Cvrček* | GF LT2 |
| [3481] | ‘Cancer pounces, snatches, latches on’ Mapping metaphor in patients’ narratives from a multilingual perspective: A corpus-based study *Jennifer Moreno* | GF LT3 |
| [6720] | Key feature analysis and multi-dimensional analysis as ways to profile disciplinary variation in journal article writing *Paul Thompson and Bethany Gray* | GF LT4 |
| [3438] | Massive Open Online Courses (MOOCs): Lexical Demands and Academic Vocabulary *Chen-Yu Liu* | CC A15 |

10.30 Coffee

|  |  |  |  |
| --- | --- | --- | --- |
| 11.00-12.30 | [9965] | **PANEL:** Methodological innovation in Corpus Linguistics: Key issues and questions *Vaclav Brezina, Raffaella Bottini, Jesse Egbert, Douglas Biber, Tove Larsson and Vaclav Cvrcek* | GF LT1 |
| 11.00 | [1934] | Exploring Sinophobic discourses during the COVID-19 pandemic: A corpus-assisted comparative discourse study of Italian, American and Australian newspapers *Ilaria Iori* | GF LT2 |
| [3451] | What does it do, to who? Defining Identity in media representations of illness *Ewan Hannaford* | GF LT3 |
| [9631] | Structure signalling in L2 undergraduate academic writing: A cross-discipline comparison *Linfeng Cai and Meilin Chen* | GF LT4 |
| [1574] | Future? Tense!: Exploring the Future in a climate change non-fiction corpus *Cinzia Bevitori and Jane Helen Johnson* | GF LT6 |
| [359] | A morphological-profiler list using AntWordProfiler: Sample comparison automated vs manual analysis. *Ana Abigahil Flores Hernandez* | CC A15 |
| 11.30 | [1231] | Xenophobic representations against China on Twitter *Cicero Soares da Silva, Aline Zamboni Milanez, Amanda Chiarelo Boldarine and Eduardo Silva* | GF LT2 |
| [8604] | A Corpus-Assisted Critical Discourse Analysis of Antidepressants in British and Chinese Newspapers and Medical Journals: the case of Escitalopram *Fang Wang* | GF LT3 |
| [9143] | Developing arguments and expressing subjectivity through modal devices in a local academic learner corpus: A small-scale study of key emergent language from a Moodle forum *Emanuela Tenca* | GF LT4 |
| [5710] | Causes, effects, and human-nature relations: A transitivity and inter-institutional analysis of UK climate change discourses *John Currie, Ben Clarke and Daniel Lees-Fryer* | GF LT6 |
| [4525] | Build your own word family: An adaptable vocabulary profiler for French, German and Spanish *Natalie Finlayson, Emma Marsden and Laurence Anthony* | CC A15 |
| 12.00 | [2453] | Subverting racist discourse through parody: A CADS analysis of A.A Gill’s essay ‘Vegetarians’ *Gill Philip* | GF LT2 |
| [7869] | Corpus Linguistics and Media Guidelines – An International Case Study on Obesity in the News *Tara Coltman-Patel, Monika Bednarek, Carly Bray and Catriona Bonfiglioli* | GF LT3 |
| [3075] | Authority and visibility in L2 (Czech) learner academic discourse: Realisations, functions and distribution of self-mention across master’s theses *Olga Dontcheva-Navratilova* | GF LT4 |
| [8025] | Examining the discourse of net-zero pledges by global corporations: A corpus-assisted analysis *Matteo Fuoli and Annika Beelitz* | GF LT6 |
| [1588] | Creating an Open-Access Thesaurus of the Welsh Language Using Natural Language Processing *Mahmoud El-Haj, Nouran Khallaf, Elin Arfon, Jonathan Morris and Dawn Knight* | CC A15 |

12.30 Lunch

|  |  |  |  |
| --- | --- | --- | --- |
| 13.30 | [5901] | Towards a systematic comparison of keyword categorisations *Nathan Dykes and Stephanie Evert* | GF LT1 |
| [6518] | Is nostalgia what it used to be? The rhetorical functions of ‘nostalgia’ and ‘nostalgic’ over time *Anna Marchi and Charlotte Taylor* | GF LT2 |
| [590] | Corpus-based discourse analysis of online health discourse *Heewon Lee* | GF LT3 |
| [8233] | Functional language in academic lectures – validating corpus linguistic findings for language assessment *Haoshan Ren* | GF LT4 |
| [4780] | The language of men's liberation: A corpus-assisted Positive Discourse Analysis *Alexandra Krendel* | GF LT6 |
| [3058] | Between Individuation and Argumentation: “A Few” vs. “Several” *Nicole Katzir* | CC A15 |
| 14.00 | [5059] | Analysing and visualising free-text comments: A corpus-based toolkit *Dawn Knight, Paul Rayson, Nouran Khallaf, Steve Morris, Mo El-Haj and Ignatius Ezeani* | GF LT1 |
| [4427] | Artificial intelligence: A corpus-assisted discourse analysis across multiple stakeholders *Jekaterina Nikitina* | GF LT2 |
| [7426] | Corpus Approach to Mental Health Discourse on Instagram: How Influencers Communicate Vulnerabilities *Aleksandra Sevastianova* | GF LT3 |
| [2806] | The effect of speaking tasks on lexical complexity and fluency: Using corpus methods to inform language assessment *Emil Tangham* | GF LT4 |
| [8355] | “The invisible disability”: Media representations and attitudes towards dyslexia and people with dyslexia in the UK and Italy *Gianmarco Vignozzi and Gloria Cappelli* | GF LT6 |
| [8345] | Super as a cross-linguistic intensifier *Chad Howe, Camila Lívio and Katherine Ireland* | CC A15 |
| 14.30 | [6088] | An AI-Assisted Annotation Approach for Non-coders: Beyond Modelling Metadiscourse *Wenwen Guan* | GF LT1 |
| [6106] | Science at Cultural Crossroads: Hedging in Arabic Discourse on Evolution *Mohammad Aboomar* | GF LT2 |
| [8221] | Negotiating the ‘good’ and ‘bad’ self: Comparing representations of people with BPD on the r/BPD and r/BPDLovedones subreddits *James Balfour* | GF LT3 |
| [8797] | Talking to an imagined interlocutor: Interactional and interpersonal features of discourse in computer-mediated semi-direct speaking assessment *Dana Gablasova, Luke Harding, Vaclav Brezina and Jamie Dunlea* | GF LT4 |
| [3002] | “O women, give charity, for verily I was shown that you make up the majority of the people of Hellfire”: Using Transitivity to explore the Islamic State through the lens of the female propagandist writer *Katie Patterson, Miguel Ángel Benitez Castro and Encarnación Hidalgo Tenorio* | GF LT6 |
| [9243] | Exploring the weakening of 'fuck' in casual conversation *Robbie Love and Anna-Brita Stenström* | CC A15 |

15.00 Afternoon tea

|  |  |  |  |
| --- | --- | --- | --- |
| 15.30 | [1736] | Metaphors, NLP and qualitative work. A metaphor annotated corpus about migration and coronavirus as a model for metaphor detection and classification.  *Dario Del Fante and Salvador Ros Muñoz* | GF LT1 |
| [5775] | Topic evolution of a political genre: Analysis and visualization of a corpus of Presidential Messages to the Congress in Chile (1832 - 2022) *Riva Quiroga* | GF LT2 |
| [8214] | The use of internal states words by bilingual children with Down syndrome: Longitudinal development, cross-linguistic variation, and register variation *Jesse Egbert, Gaëtanelle Gilquin and Tove Larsson* | GF LT3 |
| [4314] | Unveiling the Function(s) and Positions of English Discourse marker *Now* and its Persian Counterpart *Hâlâ*: A Contrastive Corpus-based Inquiry *Zahra Ghane and Mehrdad Vasheghani Farahani* | GF LT4 |
| [3963] | Exploring Saudi Novels through the Lens of Corpus Linguistics *Tareq Alfraidi* | GF LT6 |
| [4390] | "I'm so OCD lol": A corpus study of OCD as an adjective *Jordan Batchelor and Heewon Lee* | CC A15 |
| 16.00 | [2987] | Arabic Metaphor Corpus (AMC) with Semantic and Sentiment Annotation *Israa Alsiyat, Scott Piao and Mansour Almansour* | GF LT1 |
| [2467] | The prefixes anti- and pro- in Russian political discourse: A diachronic distributional semantics approach *Thomas Samuelsson* | GF LT2 |
| [7395] | A Corpus-based Study of Psycholinguistic Signs of Alzheimer’s Disease in Written Language *Xufei An and Yue Jiang* | GF LT3 |
| [8079] | A Corpus-Based Analysis of *I think* in Taiwanese EFL Learners and English Native Speakers: Research from the ICNALE Spoken Data *Ming Hung Kao* | GF LT4 |
| [3647] | “Tell me you want it, sissy”: Characterisation, narrative structure and the troubling of agency in sissy porn *Alon Lischinsky and Kat Gupta* | GF LT6 |
| [105] | How high is ‘good’, how low is ‘bad’? Do Word Embedding Spaces Encode Conventional Metaphors *Sara Bartl* | CC A15 |
| 16.30 | [1737] | A Crowdsourcing Annotated Tourism Review Corpus of Emotion *Mansour Almansour, Paul Rayson and Scott Piao* | GF LT1 |
| [9470] | A corpus-based critical discourse analysis of language ideologies in parliamentary debates about the recognition of Irish Sign Language *Robyn Cunneen and Maria Rieder* | GF LT2 |
| [4577] | Thematic role predictors in German and Russian: A corpus study *Yevheniy Skyra, Rowena Garcia and Evan Kidd* | GF LT4 |
| [3056] | ‘Windows and mirrors’ on ill health in best-selling young adult fiction *Maria Leedham, Sally Hunt and Sarah Mukherjee* | GF LT6 |
| [845] | Conceptualizing Constructions of Space Particle *Xià* in Mandarin Chinese: A Corpus Based Study *Ansley Chi-Lin Hung, Alvin Cheng-Hsien Chen and Jen-I Li* | CC A15 |

## Thursday 6th July

**9.00 Plenary: Douglas Biber, Northern Arizona University** GF LT1

|  |  |  |  |
| --- | --- | --- | --- |
| 10.00 | [7864] | Robustness of multi-dimensional register analysis with low-quality data: The case of Scottish Enlightenment writing *Aatu Liimatta, Yann Ryan, Mikko Tolonen and Tanja Säily* | GF LT2 |
| [893] | “When I first started out”: Metadiscourse in Undergraduate ePortfolios *Lydia King, Rose Delony and Jack Hardy* | GF LT3 |
| [2399] | Is multi-dimensional analysis replicable? Variables, parameters and transparency *Elen Le Foll and Vaclav Brezina* | GF LT4 |
| [2833] | A new model of variation in research articles: Optimizing multi-dimensional analysis to focus on disciplinary variation *Bethany Gray and Paul Thompson* | GF LT6 |
| [7252] | Exploring cross-linguistic representations of Web registers with a deep multilingual model *Aki-Juhani Kyröläinen, Filip Ginter and Veronika Laippala* | CC A15 |

10.30 Coffee

|  |  |  |  |
| --- | --- | --- | --- |
| 11.00-12.30 | [7509] | **PANEL:** The coronavirus infodemic: A multidimensional, discourse-based perspective *Tony Berber Sardinha, Yara Toledo Dias Romeiro, Luciana Nogueirol Lobo Marcondes, Natalia Luri Arimori Ribeiro, Marcos Oliveira, Paula Tavares Pinto, Rafael Araújo, Katherine Oliva, Tatiana Schmitz de Almeida Lopes, Deise Dutra, Mirella Whiteman, Arianne Brogini Braz, Stephan Hughes, Cicero Soares da Silva, Amanda Chiarelo Boldarine, Aline Zamboni Milanez, Eduardo Silva, Natalia Porcari Gerciano, Maria Claudia Nunes Delfino, Lucas Ferraz Escarbelin, Ana Bocorny, Simone Sarmento, Carlos Kauffmann, Renata Lamberti Spagnuolo and Mauricio Ferreira Lopes* | GF LT1 |
| 11.00 | [9149] | #ActuallyAutistic: ASD Community Building on Twitter *Savannah Brown and Jack Hardy* | GF LT2 |
| [1089] | What can a corpus tell us about workplace creativity? *Michael Handford* | GF LT3 |
| [784] | Methods of identifying culturalese in specialised discourses: A corpus investigation into the registers of Islamic law *Rana Ibrahim* | GF LT4 |
| [2874] | Multidimensional analysis of syntactic complexity development in L2 learner writing in a university IEP programme in the U.S. *Sangeun Kim and Joyce Lim* | GF LT6 |
| 11.30 | [1659] | ‘No, I can't blame that on dyslexia or anything else’—a corpus-based discourse study of the representations of dyslexia in the UK press between 2017 and 2021 *Suhua Tang* | GF LT2 |
| [8380] | A corpus analysis of creativity in care home interactions *Almut Koester* | GF LT3 |
| [6316] | US voter rights in translation: Semantic shifts in Spanish renderings of “felony” *Matt Riemland* | GF LT4 |
| [1633] | Language complexity of children’s fiction: A contrastive study of the impact of the reader, writer and language *Marketa Mala* | GF LT6 |
| [5333] | On *very few people* and *très peu de gens* – translating quantity modification with English *(a) few* and French *(un) peu* *Lobke Ghesquière and Jesse Marion* | CC A15 |
| 12.00 | [8354] | Popular and Scientific Discourse on Antidepressants in Britain and China: The case of Venlafaxine *Fang Wang, Sabine Braun and Robert Geyer* | GF LT2 |
| [1402] | “Fascinating differences”: Problem-oriented language in professional language corpora *Rachele De Felice* | GF LT3 |
| [5980] | Uncovering indigeneity in a corpus of colonial Quechua *Chad Howe* | GF LT4 |
|  | [2173] | Revealing registerial and ideological shifts in translation via word frequencies and keyword analyses: A case study of Wild Swans and Mao’s *Last Dancer* *Long Li* | CC A15 |

12.30 Lunch

|  |  |  |  |
| --- | --- | --- | --- |
| 13.30 | [3592] | Creating a Frequency Dictionary of Current British English: Methodological Considerations *Vaclav Brezina and Dana Gablasova* | GF LT1 |
| [3038] | Can low-proficiency level learners produce diverse texts? A multidimensional approach to Czech as a foreign language *Adrian Jan Zasina* | GF LT4 |
| [4540] | Discourse markers variants: A corpus study of Linguistics research article abstracts in Spanish *Hernán Robledo and Carlos Aguilar* | GF LT6 |
| 14.00 | [1747] | Which English do we need to teach to pilots? A genre analysis through corpora *Malila Prado* | GF LT1 |
| [8349] | Acquisition of Arabic Agreement System by Learners of Arabic as a Second Language *Manar Almanea* | GF LT4 |
| [9327] | A Corpus-Assisted Analysis of Shell Noun Use in Academic Lectures *Haoshan Ren and J. Elliott Casal* | GF LT6 |

**14.30 Book Launch: Encyclopedia of Shakespeare**

15.00 Closing

# Poster titles (by room order)

We will display posters throughout the duration of the conference in the room George Fox LT5. A dedicated poster session is scheduled for **Tuesday 4th July, 3.30-4.00pm**, when we request that authors are available beside their posters to discuss them with attendees.

|  |  |
| --- | --- |
| [1] | Corpus Analysis Tools: A Resource Guide *Martin Wynne* |
| [2] | Wordless: An Integrated Corpus Tool with Multilingual Support for the Study of Language, Literature, and Translation *Lei Ye* |
| [3] | Introducing CONSOS: A Spoken Corpus of New Speakers of Scots *Linda Bruce* |
| [4] | Texts of the Scottish Reformation: Creating and Normalising Corpora *Beth Beattie* |
| [5] | Analysing the Circulation of Endometriosis Terms in Different Speech Communities: Challenges for the corpus compilation *Julie Humbert-Droz* |
| [6] | Compiling and Annotating the First Electronic Learner Corpus of Romanian *Carmen Mîrzea Vasile, Ana-Maria Barbu, Enida Cincora Preda, Valentina Cojocaru, Ramona Cătălina Corbeanu, Mihaela Cristescu, Elena Irimia, Vasile Păiș and Monica Vasileanu* |
| [7] | The Saudi Novel Corpus: Progress of compilation and preliminary results *Tareq Alfraidi* |
| [8] | Machine vs human transcription: Issues in spoken learner corpus compilation *Mateus Miranda* |
| [9] | Data Protection in Corpus-based Machine Translation *Xiaojun Zhang* |
| [10] | Discourse markers, a bastion of human translation against machine translation? *Nathanaël Stilmant* |
| [11] | Evaluating the Multi-Feature Tagger of English (MFTE): Challenges and implications for corpus tool evaluations and comparisons *Elen Le Foll and Muhammad Shakir* |
| [12] | Assembling EuReCo for contrastive research: The Polish piece *Piotr Banski, Nils Diewald, Marc Kupietz and Beata Trawinski* |
| [13] | ScANT: A Small Corpus of Scene-Annotated Narrative Texts *Tarfah Alrashid and Robert Gaizauskas* |
| [15] | MultiTransRT: Multilingual Parallel Corpus and Platform for Translation Research and Teaching in China *Baorong Huang, Xiaoqian Li and Kaibao Hu* |
| [16] | Validating Terminologies and Phraseological Units Retrieved from Specialized Comparable Corpora in Lexical Semantics: An Interactive Method *Hsin-Yi Lien and Hsieh-Chih Lai* |
| [17] | A Distributional Comparison between FOLK and DeReKo *Marc Kupietz, Peter Fankhauser and Josef Ruppenhofer* |
| [18] | Analyzing French and Swedish web registers using text dispersion keywords *Saara Hellström* |
| [19] | A Spanish-English parallel corpus from financial reports *Yanco Torterolo, Sofia Roseti, Blanca Carbajo Coronado and Antonio* *Moreno-Sandoval* |
| [20] | Design of a bilingual school textbook corpus specialized in science *Miriam Buendia-Castro* |
| [21] | Good Listenership in EFL coursebooks: A context-driven Corpus Pragmatics Enquiry *Kevin Frank Gerigk* |
| [22] | Insights into pedagogical applications of corpus-based biomedical wordlists in ESP settings *Neila Cheriet* |
| [23] | Mapping shared lexical bundles onto rhetorical moves in research articles: A comparative study of paradigmatic variation *Mei Yang* |
| [24] | Causality in Spanish financial discourse *Blanca Carbajo Coronado* |
| [26] | MexLeC. A spoken and longitudinal corpus of Mexican beginner to advanced learners of English. *Ana Abigahil Flores Hernandez* |
| [27] | A Study on the Information Types of Dictionary Definition Sentences depending on the Generic Concept Categories of a Terminology: With a Focus on the Corpus of Terminology Definitions for the Physics Field in the Standard Korean Dictionary *Jun Choi, Jinsan An and Minkyu Sung* |
| [28] | A study on the usage prediction of Chinese aspect particles in Narratives using corpus and deep learning language models *Injung Jung* |
| [29] | Premodification patterns of the noun phrase in learner English writing across cohorts *Pamela Andrea Saavedra Jeldres* |
| [30] | A Corpus-based Study of the Use of Summative Connectors in Arab Leaners Academic Writing *Amani Alonayzan* |
| [31] | Investigating formulaic sequences in L2 learners’ academic writing in the EMI programmes *Tanjun Liu* |
| [33] | Intensification in Written L2 Italian in South Tyrol schools *Stefania Spina and Andrea Abel* |
| [34] | Metaphorical variation in the language of learners of English as L2: A learner corpus based study *Marcos de Oliveira* |
| [35] | Linguistic Complexity as a Dimension of Foreign Language Proficiency: Written English Performance of Czech Secondary Technical School Students *Tomáš Lorman* |
| [36] | Student Writing and Assessment Practices: Combining corpus linguistics and in-depth interviews to investigate written assessment in a distance university context *Maria Leedham, Prithvi Shrestha, Jackie Tuck and Dana Therova* |
| [37] | Playing Teacher: using corpus resources to design an intervention to support the mentoring of Early Career Teachers *Kathryn Spicksley* |
| [38] | Why Do We Use Disfluency Markers in Messaging? A Corpus-based Study on Discourse Markers in Online Messaging *Mikyung Baek, Yelin Go and Kilim Nam* |
| [39] | A corpus approach to discourses of learning in online communities *Viola Wiegand* |
| [41] | A Study of the Communist Party of China (CPC)'s Democratic Discourse (1937-1949) *Jing Wang and Nianchun Li* |
| [42] | Modelling political ideologies from parliamentary speeches *Otto Tarkka, Kimmo Elo, Filip Ginter and Veronika Laippala* |
| [43] | No straight talk here: A multi-level analysis of hedging strategies employed by the Fed Chair in press conferences *Zhipu Yang* |
| [44] | Twitter COVID-19 infodemic narratives about the Brazilian Federal Supreme Court: A lexical multidimensional analysis of the disinformation discourse *Luciana Nogueirol Lobo Marcondes, Deise Prina Dutra, Katherine Oliva Ortolani, Rafael Fonseca de Araújo and Tatiana Schimdt de Almeida Lopes* |
| [45] | A Corpus Assisted Multimodal Discourse Analysis of Public Signs in China During the Zero Covid-19 Policy: Monolingual, Visual narrative and Discursive Convergence. *Yee Chin Gan and Minjun Park* |
| [46] | “Mom is not invincible”: A corpus-based critical discourse analysis of regretting motherhood posts on Chinese Weibo *Shuning Liu* |
| [47] | Environmental Image Construction in Environmentally Sensitive Companies’ Social Responsibility Reports *Song Guo* |
| [48] | Taboo Language and Incest in the UK press (2017-2022) – Finding Absence in Corpus Linguistics *Sophie Eyssette and Gavin Brookes* |
| [49] | How British corpora reflect gender-relation through forms of address *Michael Pace-Sigge* |

# Abstracts: Panels

[5085]

***What are Fundamental Principles of Corpus Linguistics?***

Vaclav Brezina (Lancaster University) and Tony McEnery (Lancaster University).

**Abstract.** This panel will focus on a discussion about selected key Principles from a recent monograph *Fundamental Principles of Corpus Linguistics* (McEnery & Brezina, 2022, CUP). We will introduce the topic, support the discussion by references to the debate in the field of Corpus Linguistics as well as to a broader debate in the Philosophy of Science. The audience will have an opportunity to comment and ask questions. The panel will thus contribute to a field-wide reflection on what corpus linguistics is and what it does, or should do. The panel serves as an invitation formulated at the end of the book for other researchers to critique the epistemological position proposed in the volume and contrast it with their own. The panel will pay special attention to the issues of falsification and replication.

We argue that replication should not be a Cinderella of (social)science, ‘a vanishing art’ as Fresse and Peterson (2017) refer to it in the context of social science, but an essential part of (social) scientific epistemology. We will offer an empirical study demonstrating different types of replication, all aiming to maximise the potential for falsification of corpus linguistic statements. Table 1 presents an overview of three studies and includes our replication study, which will be discussed.

Table 1. Comparative overview of three studies on modals in British and American English

|  |  |  |
| --- | --- | --- |
| **Study** | **Sampling points (time)** | **Study type** |
| Leech (2003) | 1961, 1991 | Original study |
| Leech et al. (2010) | 1961, 1991 | ↑ Repetition with variation |
| Baker (2017) | 1931, 1961, 1991 2006 | ↑ Broad replication |
| McEnery and Brezina (2022) | 2006 | ↑ Direct replication |

The focus in all four studies is on the investigation of the use of modal verbs in British and American English. Modal verbs are an example of a grammatical category with a well-established pattern of variation between British and American English, i.e. an overall decline in both varieties, as indicated in the original study (Leech, 2003). Our replication study has cast a search light on specific claims made in Leech et al (2010) in particular to see whether these can be falsified or corroborated, specifically with regard to claims made about differences between British and American English.

**Keywords:** theory, philosophy of science, debate

[1200]

***Dead ends, detours, and disappointments: Looking beyond 'failed' corpus research.***

Gerlinde Mautner (WU Vienna), Mathew Gillings (WU Vienna), Paul Baker (Lancaster University), Charlotte Taylor (University of Sussex), Clyde Ancarno (King's College London) and Susan Hunston (University of Birmingham).

**Abstract.** In this panel, we aim to explore how some corpus research can lead to dead ends, result in detours, and ultimately end in disappointment. Gloomy though this may sound, the spirit in which this panel is planned is a positive and encouraging one. For as many experienced researchers do not hesitate to admit, ‘failure’ not only comes with the territory, but usually tells us a great deal about a lot of things, including the nature of the data we are dealing with, our methods, and foundational questions about epistemology.

The great majority of published work glosses over these stressors, however. The publishing process means that when we write up projects for publication, we construe the research journey in such a way that it appears entirely straightforward – as if the results simply jumped from the data. This is as much for our benefit as for our readers’, as the research story must be as easy as possible to understand. Yet by simply ignoring complexity, we may be doing ourselves and our field a disservice. Junior scholars may be fooled into thinking that the process is mechanical, and feel discouraged when their data doesn’t play ball. What is more, they may quite rightly feel that being open about failure, and dealing with it, is a privilege that is unique to tenured faculty. Yet the latter too are subject to institutional and reputational pressures.

In this panel, we invite four established researchers to draw upon their own experiences and reflect upon projects of theirs that struggled to make progress. First, Paul Baker will discuss issues that he has personally encountered: projects which had to change direction half-way through; data which turned out to be faulty and threw up strange results; cases where the research simply resulted in uninteresting results; and challenges with applying old methods to new questions. He will offer some practical guidance on dealing with such dead-ends which might, in some cases, simply be to let go altogether. Charlotte Taylor will similarly draw on her personal research experience to discuss the ‘detours’ of research journeys. This is a space to think about the indirect progress of research projects, which is often in sharp contrast to the linear image given in published pieces coming out of them. Afterwards, Clyde Ancarno will focus on issues encountered in the study of Yoruba and Gambian languages, caused by, for example, the prevalence of code-switching, and inconsistencies in the way words/phrases are truncated and spelled. Finally, Susan Hunston will discuss the essentially comparative nature of most corpus linguistic work: the issues with establishing ‘baseline’ corpora, and the problem that researchers are faced with when they simply don’t have two corpora available to compare.

We hope that the panel will kick-start a renewed sense of transparency and honesty in the field (ironically, an honesty that is often found in PhD theses, but not published work). As Baker (2018: 289) writes, ‘nobody talks about “failed corpus research” […] I wonder if we could collectively learn as much from our mistakes as our successes and perhaps we should worry less about our images.’

**Schedule**

10 min – Introduction

10 min – Paul Baker

10 min – Charlotte Taylor

10 min – Clyde Ancarno

10 min – Susan Hunston

30 min – Discussion

**Keywords:** epistemology, methodology, reflexivity

[5327]

***Corpus linguistic approaches to school-aged children’s writing.***

Philip Durrant (University of Exeter), Hildegunn Dirdal (University of Oslo), Yaling Hsiao (Unversity of Oxford) and Victorina Gonzalez-Diaz (University of Liverpool).

**Abstract.** Learning to write effectively is one of the central tasks of education, and modelling the writing of school-aged children is one of the key contributions that applied linguists can make to educational research. There has, accordingly, been a long history of research studying key linguistic features of children’s writing. Much of this work can be described as corpus-like in that it has collected samples of texts and systematically studied linguistic features both quantitatively and qualitatively. However, most has been based on ad-hoc collections of writing, elicited by researchers for the purposes of a single study. Very few have shared the corpus linguistic aims of creating a representative body of authentic writing or of making texts available to the wider research community.

In recent years however, this picture has started to change, with several projects sowing the seeds of a distinctively corpus linguistic strain of child writing research. This panel aims both to highlight this work and to further research by bringing together corpus linguists working on different types of child writing and within different theoretical perspectives.

Specifically, the panel will discuss recent and ongoing work on both school writing and extra-curricular writing from the varied perspectives of first language education (Durrant), foreign language education (Dirdal), psychology (Hsiao) and historical linguistics (Gonzalez-Diaz). After a general introduction, setting out the aims of the session, each contributor will speak on their own work for 15 minutes. In each case, they will discuss:

* the main aims of their research;
* the corpus/corpora that they work with;
* the corpus linguistic methods they use;
* key findings from their research;
* their plans for future work.

Within the above rubric:

Philip Durrant (University of Exeter) will discuss his work on how school children’s writing develops through the course of their educational careers in England, with reference to the Growth in Grammar corpus.

Hildegunn Dirdal (University of Oslo) will discuss her work on complexity development and the use of different clause types in the writing of Norwegian school children, using data collected for the TRAWL Corpus and the MULTIWRITE project, which focuses on L1, L2 and L3 connections and comparisons.

Yaling Hsiao (University of Oxford) will discuss her work on lexical and syntactic complexity in school-aged children's narrative writing using a large developmental corpus derived from the national BBC 500 Words competition.

Victorina Gonzalez-Diaz (University of Liverpool) will discuss the Writing over Time project, a collaborative corpus-based project which investigates variation and change in primary school children’s writing before and after the implementation of the National Curriculum.

We will then open the floor to general discussion. All participants will be invited to reflect on the strengths and limitations of current work, on points of overlap and divergence between projects, and on prospects and priorities for future research and collaboration.

**Keywords:** child language, school language, academic language, second language, third language, historical linguistics, complexity

[6712]

***Discourse units and sub-register variation in British English conversation.***

Jesse Egbert (Northern Arizona University), Douglas Biber (Northern Arizona University) and Tony McEnery (Lancaster University).

**Abstract.** Previous research on conversational language has focused either at the micro-level (characteristics of turns and adjacency pairs; e.g. Rühlemann, 2007; Stenström, 2014) or the macro-level (characteristics of entire recorded conversations, often in comparison with other registers; e.g. Biber et al., 2021). We adopt a third approach which is focused on identifying and describing functional discourse units (DUs) within conversations and describing the linguistic characteristics of those units. Then, conversational sub-registers can be identified and described, consisting of sets of DUs with shared functions. The purpose of this panel is to introduce the methods we use to accomplish this and the findings we have produced to date. The panel is organized into three parts:

1. Conversational discourse units: Segmentation and annotation;
2. Conversational sub-registers: Classification of DUs; distribution and linguistic description of sub-registers
3. Discourse units and conversational narrative

In Part 1, we introduce the construct of a conversational DU, and describe the ways in which extended conversational interactions can be described as sequences of DUs. We describe and illustrate the ways in which DUs are self-contained and are internally coherent for their communicative function. Building on this theoretical framework, we describe the methods we used to segment 479 extended conversations in the BNC Spoken 2014 corpus into coherent DUs and annotate each DU for the degree to which it relies on nine communicative functions (e.g., describing past-time events, situation-dependent commentary). Each communicative purpose is introduced and illustrated. In addition, we describe how DUs often combine multiple communicative purposes that differ in their importance for the particular interaction.

In Part 2, we begin by discussing what the notion of register means within the discourse domain of conversation. Because there is no agreement on the set of sub-registers found in conversation, we introduce our bottom-up method for identifying conversational sub-registers. This was done by attempting to classify each DU into a single conversational sub-register, or a hybrid sub-register category, based on shared communicative functions of the DUs. We document the ways in which different communicative purposes tend to co-occur in DUs and the prevalence of each sub-register and hybrid sub-register in the BNC Spoken 2014. We then describe linguistic similarities and differences among the sub-register categories using multi-dimensional scaling applied to a large set of lexico-grammatical features.

In Part 3, we explore to what extent the discourse unit approach can provide insight into or present a challenge to established units of analysis in discourse. Using the example of the narrative, a well established metastructure in discourse analysis, we will explore the extent to which a discourse unit perspective on narrative is coherent with established approaches to narrative. The focus here will be on the Labovian model of narrative, and the goal will be to see whether the two analytical schemes can be harmonised and, if so, what that tells us about both schemes.

We conclude the panel with a Q+A session, allowing audience members to ask questions and provide feedback about the methods and results of the project to date.

**Keywords:** conversation, interactive speech, discourse units, communicative purpose, discourse function, register variation, narrative

[9965]

***Methodological innovation in Corpus Linguistics: Key issues and questions.***

Vaclav Brezina (Lancaster University), Raffaella Bottini (Lancaster University), Jesse Egbert (Northern Arizona University), Douglas Biber (Northern Arizona University), Tove Larsson (Northern Arizona University) and Vaclav Cvrcek (Charles University, Prague).

**Abstract.** This panel follows from a successful series of Lancaster symposia on the innovation in Corpus linguistics, which offer multiple short contributions on key current issues in the area of corpus design and methodology. The focus of the panel will be on innovation in the field and the vision for the future. Each of the presenters will briefly introduce their topic, which will then be discussed by the panel. The audience will have an opportunity to comment and ask questions.

The topics discussed will include: i) collocations, ii) grammatical annotation, iii) latent variables, iv) key feature analysis, v) lexical sophistication.

Collocations.

Two methods to harvest non-adjacent associations (or network of concepts) will be suggested, which complement traditional collocation analysis, which identifies associations of concepts in the discourse: Market Basket Analysis (Cvrček & Fidler, 2022) and Companions (Cvrček & Fidler, forthcoming). The first method takes a number of keyword lists (one per each text in the corpus) to determine how likely a keyword co-occurs with other keywords in the same text. The second method allows measuring how two words, triggered by some real-life event, start occurring at the same time and have similar (coinciding) frequency developments.

Grammatical annotation of corpora.

One of the most important linguistic assumptions in any part-of-speech and lemmatized corpus is that the grammatical categories have been assigned accurately and consistently. Reported accuracy statistics for English of existing tagging systems such as Tree Tagger, CLAWS, spaCy are high (96-98%); however, their performance is not consistent across genres/registers and individual lexical items (cf. Evert, 2009). The implications for corpus linguistic analysis will be discussed.

Author characteristics as latent variables.

We can arguably do more to collect and use information about the author or speaker whose production is in our corpus. For example, researchers looking at L2 English may want to learn more about the learners’ aptitude and motivation for learning a language and how these affect some dependent variable. Variables of this kind are – and should be modeled as – latent variables. Latent variables are variables that lack directly observed (or, oftentimes, observable) realizations and that therefore need to be inferred through variables that can be observed (e.g., Bollen, 2002; Larsson, Plonsky, & Hancock, 2022).

Key feature analysis.

To date, corpus-based methods for comprehensive comparisons of language varieties have fallen into one of two camps: (1) MD analysis—a complicated multivariate approach based on analysis of functionally motivated linguistic features in each text of a corpus, and (2) keyword analysis—a simple method for identifying words with statistically-skewed distributions in one corpus when compared with another. We introduce a complementary technique—key feature analysis—a simple quantitative method that compares the texts in two varieties with respect to a set of functionally motivated lexico-grammatical features. Key feature analysis is a powerful new corpus-based technique that combines the simplicity of keyword analysis with the flexibility and text focus of MD analysis.

Lexical sophistication.

Measures of lexical sophistication based on frequency depend on (i) the lexical unit to identify words, (ii) the reference corpus to measure frequency, (iii) the method that rates frequency in a target text, and (iv) operational choices to process the data (e.g., Bottini, 2022; Bottini, & LeFoll, in prep.; Kyle, 2019). However, these methodological details are often underreported in the field, raising questions about the validity and reliability of lexical sophistication measures. Possible future directions are discussed.

**Keywords:** methodology, innovation, critical reflection

[7509]

***The coronavirus infodemic: a multidimensional, discourse-based perspective.***

Tony Berber Sardinha (Pontifical Catholic University of Sao Paulo), Yara Toledo Dias Romeiro (Pontifical Catholic University of Sao Paulo), Luciana Nogueirol Lobo Marcondes (Sao Paulo Technical College), Natalia Luri Arimori Ribeiro (Pontifical Catholic University of Sao Paulo), Marcos Oliveira (Pontifical Catholic University of Sao Paulo), Paula Tavares Pinto (Sao Paulo State University), Rafael Araújo (Sao Paulo Technical College), Katherine Oliva (Pontifical Catholic University of Sao Paulo), Tatiana Schmitz de Almeida Lopes (Sao Paulo Technical College), Deise Dutra (Minas Gerais Federal University), Mirella Whiteman (Pontifical Catholic University of Sao Paulo), Arianne Brogini Braz (Pontifical Catholic University of Sao Paulo), Stephan Hughes (Pontifical Catholic University of Sao Paulo), Cicero Soares da Silva (Pontifical Catholic University of Sao Paulo), Amanda Chiarelo Boldarine (Pontifical Catholic University of Sao Paulo), Aline Zamboni Milanez (Pontifical Catholic University of Sao Paulo), Eduardo Silva (Goias State University), Natalia Porcari Gerciano (Pontifical Catholic University of Sao Paulo), Maria Claudia Nunes Delfino (Sao Paulo Technical College), Lucas Ferraz Escarbelin (Pontifical Catholic University of Sao Paulo), Ana Bocorny (Rio Grande do Sul Federal University), Simone Sarmento (Rio Grande do Sul Federal University), Carlos Kauffmann (Pontifical Catholic University of Sao Paulo), Renata Lamberti Spagnuolo (Pontifical Catholic University of Sao Paulo) and Mauricio Ferreira Lopes (Pontifical Catholic University of Sao Paulo).

**Abstract.** Although at this time the COVID-19 pandemic has receded, its accompanying infodemic has not -- in fact the infodemic might be one of most enduring and sinister legacies of the pandemic. An infodemic is generally understood as the rapid spread of large quantities of untrustworthy information, whose dissemination is made possible by digital communication, particularly social media. In this panel, we see the infodemic as an ecosystem of discourses, that is, 'ways of seeing the world, of constructing objects and concepts in certain ways, of representing reality' (Baker & McEnery, 2015, p. 5), which interconnect, interact and feed off each other, creating a coherent rather than a chaotic worldview.

Methodologically, we rely on Lexical Multidimensional (LMD) Analysis to detect the different discourses. In short, firstly, large corpora of tweets and academic articles in either Brazilian Portuguese or English were scraped, tagged for POS and lemmatized. Next, lemma counts were tabulated, and entered in a factor analysis. Finally, the factors were interpreted as underlying discourses. For the visual analysis, first each image was automatically labeled using Google Cloud Vision. Next, the label counts were entered in a factor analysis, which yielded the sets of correlated visual features in the image corpus. Finally, these factors were interpreted as visual dimensions, corresponding to the major visual patterns present in the corpus.

Most of the individual contributions trained a machine learning model to detect infodemic tweets. Labeling involved several rounds of annotation, and both intra- and interrater agreement reached at least .7. The models were trained and tested, resulting in accuracy rates of up to 90%.

Contributions in the panel will address different aspects of the infodemic in social media: (1) The discourses around ‘genocide’, a term characterizing the death toll caused by the pandemic. This paper will present the major lexical dimensions associated with ‘genocide’-blaming discourses used on both sides of the political spectrum. (2) The discourses around the (Brazilian) Supreme Court handling of the pandemic, which came under heavy criticism from right-wing groups for allegedly promoting censorship of social media and restricting individual freedoms. (3) Science denying discourses supporting the antivax movement, which saw a resurgence in Brazil and the US – the dimensions include disseminating side-effects and relying on pseudo-science for delegitimization. (4) The xenophobic representations of China, which reflect not only its leading role in the pandemic (as a source of the virus, as a worldwide exporter of vaccines) but an opposition against its growing geopolitical and economic importance. (5) Representations of mental disorders during the pandemic. Despite the rise in mental illness cases during the pandemic, the discourses around mental disorders are conflicting, downplaying their importance. (6) Representations of fake news. This paper will look at how the term fake news has been employed on Twitter. The dimensions reflect different discursive strategies, such as accusing and deflecting criticism. (7) Discourses in pseudo-science academic articles. This contribution will present the dimensions around the discourses used by academic pseudo-science articles scraped from a large repository.

**Keywords:** Infodemic, COVID-19, Lexical Multidimensional Analysis, Discourse

# Abstracts: Papers

[105] ***How high is ‘good’, how low is ‘bad’? Do Word Embedding Spaces Encode Conventional Metaphors.***

Sara Bartl (University of Birmingham).

**Abstract.** Word embeddings represent words as vectors in a multi-dimensional space, whereby a word’s vector is determined by its co-occurrence patterns with other words in the corpus. The more likely words are to occur in similar contexts in that corpus, the closer they are thought to be in meaning and as a result, they will be closer together in the embedding space (Lenci, 2018). While word embeddings have greatly improved the performance of automatic metaphor identification algorithms (Tong, Shutova and Lewis, 2021), it seems that the quest to improve benchmark scores has blindsided research a fundamental question. Namely, whether and to what extent the underlying large-scale co-occurrence patterns of word embeddings already capture figurative meaning.

In this paper, I address this by extending Grand et al.’s (2022) semantic projection method. Semantic projection is a method that aims to recover relationships between words through vector projection. This is done by establishing a spectrum between antonyms (e.g. ‘up’ and ‘down’) and then ‘projecting’ words onto that spectrum. This method reveals how words relate to each other along a given spectrum (e.g. direction). As a case study, this method is applied to vertical orientation metaphors of emotional valence (UP IS GOOD/DOWN IS BAD) and quantity (MORE IS UP/LESS IS DOWN). A growing body of behavioural and text-based research attests a figurative relationship between vertical orientation (up/down) and emotional valence (good/bad) and quantity (more/less) (e.g. Deignan, 2005; Woodin and Winter, 2018). If word embedding spaces encode figurative meaning, an application of the semantic projection method should reveal that a lexical item like ‘good’ is ‘higher’ on the up/down spectrum than a lexical item such as ‘bad’.

The results show that while the locations of some target domain items appear idiosyncratic, word embeddings based on co-occurrence patterns do encode the structure of the metaphors examined.

**Keywords:** word embeddings, metaphor, computational linguistics, semantic projection

[128] ***Web corpora for under-resourced languages.***

Vit Suchomel (Lexical Computing), Miloš Jakubíček (Lexical Computing) and Ondřej Matuška (Lexical Computing).

**Abstract.** Text corpora provide essential information for studying natural language phenomena, language learning, lexicography and for building language models. In the case of under-resourced languages, one has to make extra effort to find as many texts as possible to build corpora large enough for intended purposes. In this paper, we share our experience with compilation of corpora from the web in 25 under-resourced languages. Challenges we had to face in the process and opportunities of providing resources in languages that have recently increased in demand are discussed.

Since 2017, we have crawled top level domains of countries where under-resourced languages are spoken. To compile a corpus, text was extracted from web pages, then boilerplate, duplicates and non-text were removed. As for European languages, we have e.g. over 140 million Irish tokens, 500 million Albanian tokens and 90 million Maltese tokens. Another set of corpora in Asian languages was built for a lexicography project to deliver 45,000 headword dictionaries based on 120 million Lao tokens and 230 million Tagalog tokens (and other collections). In that work, we found how many tokens is enough for a mid-sized bilingual dictionary. We have also experimented with African languages native to Ethiopia and Nigeria, getting e.g. over 30 million Amharic tokens.

Although we have to deal with issues of languages less represented on the web such as a low genre and topic variety of sources and the fact that smartphones arguably lead to communication through visual media rather than text in some countries, building corpora in under-resourced languages provide many opportunities: There is no language research without a corpus. Language data is necessary for smartphone applications such as predictive writing. Finally, a modern dictionary enables the preservation of endangered languages and standardisation of languages spoken but not much used for written communication.

**Keywords:** under-resourced languages, web corpora, lexicography

[154] ***Network of Grammar Lexemes in 16th- and 17th-Century English Grammar Writing.***

Beatrix Busse (University of Cologne), Nina Dumrukcic (University of Cologne), Sophie Du Bois (University of Cologne) and Ingo Kleiber (University of Cologne).

**Abstract.** Most terms that originally referred to fields of study, “have also come to denote the particular field of language itself” (Fenn 2022, 23), blurring the lines between grammar and linguistics. Before the introduction of approaches such as generative grammar (Chomsky 1957) and systemic functional grammar (Halliday 1994) the earliest stages of English grammar writing relied on the categorization of Latin grammar (Algeo 1985). However, in the late 17th century, the English language became more standardized and extended to domains of language use, where the classical languages had previously reigned (Nevalainen 2006, 42). This and other socio-political events invoked a turn towards English being acknowledged as a subject of study in itself (Beal 2004, 102). The scope of grammar has changed from including orthography and prosody, to focusing on morphology and syntax (Walmsley 1999, 2495). This study investigates the diachronic shift in terminology employed for grammatical fields of study and their related concepts. A list of terms known to be used for fields of study that fall under ‘grammar’ is extracted from the Historical Thesaurus of English (Kay et al. 2023). For example, Morphology only emerges as a term in 1869–, and was previously referred to as Etymology c1475– and Wordlore 1840–.

The corpus of 16th- and 17th-century English grammars associated with the HeidelGram project (see http://heidelgram.de) and the list of search terms allow us to explore semantic fields and shifts by means of concordance and collocational analysis. We will generate networks of the different structures that occur in grammars during this period enabling us to compare whether there have been notable shifts in what grammatical categories the authors discuss. Our aim is to detect traces of modern theoretical approaches such as Lexical-Functional Grammar (see Bresnan et al. 2016) in the earliest grammars and how they evolved over time.

**References:**

Algeo, John. 1985. “The Earliest English Grammars” in Historical and Editorial Studies in Medieval and Early Modern English: For Johan Gerritsen. Edited by Mary-Jo Arn and Hanneke Wirtjes, 191–207. Groningen: Wolters-Noordhoff.

Beal, Joan C. 2004. English in Modern Times. London: Arnold.

Bresnan, Joan, Ash Asudeh, Ida Toivonen, and Stephen Wechsler. 2016. Lexical-Functional Syntax. Second edition. Chichester: Wiley Blackwell.

Chomsky, Noam. 1957. Syntactic Structures. The Hague/Paris: Mouton.

Fenn, Peter. 2022. A Student's Advanced Grammar of English (SAGE). 2., revised edition. Stuttgart: utb GmbH.

Halliday, M.A.K. 1994. Introduction to Functional Grammar, Second edition. London: Edward Arnold.

Kay, Christian, Marc Alexander, Fraser Dallachy, Jane Roberts, Michael Samuels, and Irené Wotherspoon (eds.). 2023. The Historical Thesaurus of English (2nd edn., version 5.0). University of Glasgow. https://ht.ac.uk/.

Michael, Ian. 1970. English Grammatical Categories: And the Tradition to 1800. Cambridge: Cambridge University Press.

Nevalainen, Terttu. 2006. An Introduction to Early Modern English. Edinburgh: Edinburgh University Press.

Walmsley, John. 1999. “English grammatical terminology from the 16th century to the present” in 2. Halbband: Ein internationales Handbuch zur Fachsprachenforschung und Terminologiewissenschaft. Edited by Lothar Hoffmann, Hartwig Kalverkämper, Herbert Ernst Wiegand, Christian Galinski, and Werner Hüllen, 2494-2502. Berlin/New York: Mouton De Gruyter.

**Keywords:** historical corpus, English grammar writing, grammatical phenomena, network analysis, seventeenth century

[322] ***Discursive construction of key issues in Australian citizenship legislative enquiry from 2007 to 2017: A corpus-assisted discourse analysis.***

Remart Dumlao (Monash University).

**Abstract.** This study explores the use of public submission in Australia to elicit information, feedback, and narratives on policies, particularly citizenship policies. It examines the limited understanding of how public submission is utilized to explore critical issues and expand our understanding of legislated policies. The research focuses on two significant periods of Australian citizenship policy: the 2007 bill comprising 708,769 words and the proposed 2017 amendment consisting of 547,901 words. Through collocation and concordance analysis, the study identifies frequent topics produced by Odd ratios such as evaluative attribution of aspiring citizens, and Australian ideologies. Employing the Discourse Historical Approach to critical discourse analysis, the study investigates the discursive strategies of nomination and predication used in these public submissions. The findings reveal racialized attributes and competence-based evaluations associated with aspiring citizens (including refugees, migrants, and dependent visa holders). This study shows how Critical Approaches to Critical Discourse Analysis can be used to analyse public comments. It also shows how important it is, from a methodological point of view, to look at civic issues and the role of public comments in how the government talks to the public.

**Keywords:** Citizenship, CADS, Public submission, government communication

[359] ***A morphological-profiler list using AntWordProfiler: Sample comparison automated vs manual analysis.***

Ana Abigahil Flores Hernandez (Universidad Autonoma de Queretaro).

**Abstract.** The LCR tools development field has dedicated little attention to complex words and the derivational morphemes in them, even though they play an important part in learners lexicon development (Arslan et al, 2020; Fontana, 2013). Some of the few studies dedicated to morphemes as Fuchs and Werner (2018), Ardebol (2012), Khor (2012) and Fontana (2013) among others, have made use of a semi-automated analysis using a tagged word list and then selecting suffixes (manually) or even making a whole manual ´pen and paper´ analysis.

This work introduces a proposal for an automated analysis of derivative suffixes in a learner corpus of written essays of 25,642 tokens and four proficiency levels. Here it is reported a sample comparison of a manual and automatic search using a sample list of 80 suffixes and its reference derivatives taken form BNC. Results show similar raw token counting between automated and manual analysis of these two suffixes in all level sections of the corpus with minimum differences.

The final suffixes list is expected to cover a total of 153 suffixes adapted from Stein (2007) and the OED (2019) and a derivatives reference list of 17,943 word types (Laws and Ryder, 2019). No doubt, the creation of this expected extensive list may be a powerful tool for complex words analysis that can solve the problem of morpheme variants and morphemes sharing the same form, since it is not based on word ending recognition. On the other hand, some limitations may be the identification of suffixes in misspelling words (too frequent in learner interlanguage) and the number of reference lists the software can hold for processing at a time.

**References**

Anthony, L. (2021). AntWordProfiler (Version 1.5.1) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Ardébol, J. (2012). A Morpheme Study in a Corpus of Secondary School EFL: The “– (e)s” Morpheme for 3rd Person, Plural and Possessive. MA dissertation in Teacher Training. Granada: University of Granada.

Arslan, F., Mahmood, M. and Rasool, A. (2020) A Corpus-Based Comparative Study of Derivational Morphemes Across ENL, ESL, EFL Learners Through ICNALE. Linguistic Forum: 2, 1-12.

Stein, G. (2007). A Dictionary of English Affixes: Their Function and Meaning. Munich: Lincom Europa.

Fontana, U. (2013) A morpheme order study based on an EFL learner corpus: A focus on the dual mechanism. MA dissertation in Teacher Training. Granada: University of Granada.

Fuchs, R. and Werner, V. (2018) The use of stative progressives by school-aged learners of English and the importance of the variable context. International Journal of Learner Corpus Research 4:2, 195-224.

Khor, S. (2012). A corpus based study in morpheme acquisition order of young

learners. A comparison of Swedish students in grade 6 (Tesis doctoral)

Laws, J.V. y Ryder C. (2017). MorphoQuantics: A corpus of derivational morphology in adult spoken. English. Available from: <http://morphoquantics.co.uk> [31/05/2019]

Meurers, D. (2015) Learner Corpora and Natural Language Processing. In Granger, Sylviane, Gilquin, Gaëtanelle & Meunier, Fanny, eds., The Cambridge Handbook of learner corpus research. Cambridge: Cambridge University Press.

OED (2019). Oxford English Dictionary – OED online. Oxford: Oxford University Press. Available from: <http://www.oed.com>

**Keywords:** Learner Corpora, Derivational Morphology, Corpus Tools

[362] ***English binominal noun phrases as seen through their Czech counterparts.***

Gabriela Brůhová (Charles University) and Kateřina Vašků (Charles University).

**Abstract.** Based on the data from a parallel English-Czech corpus, the present paper offers an analysis of 400 English binominal noun phrases (BNPs) through their Czech translation correspondences. In English BNP (e.g. a giant of man, a hell of a problem), “the first noun [N1] denotes a property or quality that is predicated of the referent of the second noun [N2]” (Aarts 1998: 118). It has been shown that it is N2 that functions as the semantic head, while of-N1 functions as a modifier (Aarts 1998; Kim & Sells 2015).

This paper has a twofold aim: a) to identify and describe the Czech means which perform the same functions as the English BNP, b) to further specify the functions of the construction as seen through the Czech counterparts. Based on ten Wolde’s account of BNPs (2022) we assume that BNP can be subclassified into several types (e.g. evaluative, intensifying) which can be revealed through the translational counterpart.

Our preliminary results show that the relation between N1 and N2 is often expressed by an modifier + head noun in Czech (ex. 1). Other possible correspondences are synthetic counterparts (ex. 2), or a more explicit construction with a relative clause (ex. 3).

(1) that fool of a girl – ta pitomá holka ‘that stupid girl’

(2) one hell of a brain – kapacita ‘expert’

(3) a bleak cage of a vestibule - chodbička, která vypadala jako ponurá klec ‘a vestibule that looked like a bleak cage’

**References:**

Aarts, B. (1998). Binominal noun phrases in English. Transactions of the philological society, 96(1), 117-158.

Kim, J. B., & Sells, P. (2015). English binominal NPs: A construction-based perspective1. Journal of linguistics, 51(1), 41-73.

ten Wolde, E. (2022). From an icy hell of a night to a hell of a fine story. English Noun Phrases from a Functional-Cognitive Perspective: Current issues, 221, 279.

**Keywords:** Binominal noun phrases, Czech, contrastive approach

[387] ***Textual Functions and Aspects of Underspecification of Discourse Markers in Translation: A Parallel Corpus-based Investigation of English and Persian.***

Zahra Ghane (University Düsseldorf) and Mehrdaad Vasheghani Farahani (Leipzig University).

**Abstract.** As the elements which are polyfunctional and used to construct and mediate addressor- addressee interactions, discourse markers have always been an intriguing line of research in spoken and written discourse. Taken this into consideration, this study was an effort to investigate, quantitatively and qualitatively the function(s), underspecification and translation(s) of five most prevalent discourse markers in an English-Persian parallel corpus. To this end, a fine-grained parallel corpus of law and political texts was created to analyze the designated discourse markers. The analysis of the parallel concordance lines demonstrated that functional translation of the metadiscourse features was the leading translation strategy in Persian language. Apart from that, there were degrees of underspecifications including omission and extension of metadiscourse features. However, the analysis of parallel concordance lines illustrated that the function(s) of discourse markers did not change from the source language into the target language, as they were found to be context dependent. The results of this study can have implications for semantics, pragmatics, translation studies and corpus linguistics.

**Keywords:** Parallel Corpora, Semantic Underspecification, Translation, Functions of Language, English and Persian Languages

[567] ***Creating and analysing a multimodal corpus of obesity news texts with Google Cloud Vision’s automatic image tagger.***

Luke Collins (Lancaster University) and Paul Baker (Lancaster University).

**Abstract.** Media representations contribute to public and professional understandings of important health issues, like obesity (Lyons, 2000) and visual representations of people with obesity presented in news media have been shown to be stigmatising (Heuer et al., 2011). Large-scale analyses of media discourses using corpus approaches have typically investigated only the written text of news articles, due to the challenges of documenting image (and other visual) content in a machine-readable format. We describe the creation and analysis of a small multimodal corpus of UK news articles about obesity, where tags were assigned to images in the articles using the automatic tagger Google Cloud *Vision*. Our analysis comprises two stages: i) the comparison of tags derived from images in UK news articles compared with those generated from the World Obesity Federation (WOF) image bank, which was curated to support journalists in avoiding stigmatising and stereotypical portrayals of individuals with obesity; ii) the exploration of text-image relations in obesity news articles through keyness and collocation analysis.

We find only minimal consistencies between images appearing in news articles and images developed by the WOF. We observe differing representations across news publications, with some focussing on body positive representations while others offer more dehumanising representations of pinched (faceless) abdomens. The WOF image bank favours a position of ‘body neutrality’, which was manifest in visual representations of a greater range of activities, including how people interact with food.

Through this application, we consider the potential of the automated image annotation tool for conducting large-scale analyses of image content and media representations as part of a wider corpus linguistic approach.

**References**

Lyons, A. C. (2000). Examining media representations: Benefits for health psychology. Journal of Health Psychology, 5(3), 349-358. <https://doi.org/10.1177/135910530000500307>

Heuer, C. A., McClure, K. J. and Puhl, R. M. (2011). Obesity stigma in online news: a visual content analysis. Journal of Health Communication, 16(9), 976-987. <https://doi.org/10.1080/10810730.2011.561915>

**Keywords:** obesity, image representations, multimodality, health discourse

[590] ***Corpus-based discourse analysis of online health discourse.***

Heewon Lee (Georgia State University).

**Abstract.** Digital media on the internet, such as online discussion forums, are increasingly used by healthcare consumers to reach out to virtual communities to seek advice for managing their conditions (Miller & Pole, 2009). Thus, online discussion forums provide rich descriptions of what sufferers of mental health conditions experience. Recent research has shown corpus linguistic methods to be a valuable way of investigating authentic language use by those affected with mental health disorders (e.g., Hunt & Brookes, 2020). This study sought to learn more about bipolar disorder from a lived experience perspective, focusing on how people who have bipolar disorder relate to the disorder and related symptoms. A corpus of posts made to an online forum serving as an anonymous community for those impacted by bipolar disorder was collected and cleaned using Python. Then, a frequency list of lexical words (e.g., bipolar, mania) and 3 and 4-grams (e.g., a manic episode, was diagnosed with) related to bipolar disorder were generated. Next, the action and state verbs (e.g., BE, DIAGNOSED, GO INTO) frequently used in conjunction with those phrases related to bipolar disorder (i.e., bipolar, mania, hypomania, depression, manic episode, depressive episode) were investigated. Concordance lines, including these phrases, were investigated. Findings gathered from these analyses suggest that people who experience bipolar disorder have different ways of verbalizing their relationship to their disorder: some identify themselves with the disorder (i.e., I am bipolar), some put distance between themselves and the disorder (i.e., I am diagnosed with bipolar disorder), whereas others express their experience with the disorder through spatial metaphors (i.e., manic episode is coming). The results have implications for raising awareness about how mental health consumers conceptualize and verbalize their relationship with mental health conditions and related symptoms.

**Keywords:** Corpus linguistics, health discourse, online discourse

[618] ***Building blocks of fairy tale openings and closings in the TransGrimm Corpus.***

Christina Sanchez-Stockhammer (TU Chemnitz) and Asya Yurchenko (TU Chemnitz).

**Abstract.** The fairy tale is famously fluid and nomadic (Saego 2018: 305), as it can simultaneously exist in multiple versions not only in its original language but also in translations. With this flexibility of the genre in mind, it is the goal of the TransGrimm corpus project to investigate the original German editions of the Grimms’ “Kinder- und Hausmärchen” and their English-language translations from linguistic perspectives using quantitative, corpus-based approaches (e.g. Sanchez-Stockhammer 2020).

In the paper proposed here we focused on fairy tale openings and closings as represented by the 1857 German edition and its 1884 English translation by Margaret Hunt. The goal was to find patterns in the way the Grimms began and concluded their fairy tales and to determine how these formulae, along with references to time and the presence of archetypal characters, work together to “invoke the presence of the Taleworld” (Nunes & Greenhill 2018: 26) in the two versions.

Our results show, among other things, that the vast majority of the time references in the openings are of a generic and vague nature (e.g. “a long time ago”). Contrary to popular belief, “once upon a time” is not the most common English fairy tale opening, as “once on a time” is almost twice as frequent. The majority of the first-named entities are generic and archetypal, with professions (king, miller, servant) as by far the most frequent category and a clear underrepresentation of proper names. Cluster analyses of the collocates of the verb “lived” in the closings revealed ‘building blocks’ of fairy tale endings, which occur individually or in various constellations of FEELING + TIME + COMPANY + QUALITY (e.g., “lived [happily]FEELING [together]COMPANY [until their death]TIME”).

All this contributes to the ‘operationalized genericity’ of the fairy tale genre (Aldred 2018: 264) and serves to invoke its characteristic Taleworld(s).

**References**

Aldred, B. Grantham. 2018. Language. In P. Greenhill, J. T. Rudy, N. Hamer & L. Bosc (eds.), 263-271. The Routledge companion to media and fairy-tale cultures. London: Routledge.

Grimm, J., & W. Grimm. 1857. Kinder- und Hausmärchen. 7th edn. Göttingen: Dieterichsche Buchhandlung.

Grimm, J., & W. Grimm. 1884. Grimm’s household tales (M. R. Hunt, Trans.). London: Bell and Sons.

Nunes, V. & P. Greenhill. 2018. Constructing fairy-tale media forms: Texts, textures, contexts. In P. Greenhill, J. T. Rudy, N. Hamer & L. Bosc (eds.), The Routledge companion to media and fairy-tale cultures, 20-28. London: Routledge.

Seago, K. 2018. Translation. In P. Greenhill, J. T. Rudy, N. Hamer & L. Bosc (eds.), The Routledge companion to media and fairy-tale cultures, 299-308. London: Routledge.

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, 305-327. Bern: Peter Lang.

**Keywords:** fairy tale, genre, TransGrimm corpus, opening sentence, closing sentence, building blocks, collocations, English, German, time references

[784] ***Methods of identifying culturalese in specialised discourses: A corpus investigation into the registers of Islamic law.***

Rana Ibrahim (Dublin City University (DCU)).

**Abstract.** This research sets out to investigate how culture-specific or signature concepts are rendered in English-language discourse on Islamic, or ‘sharia’ law, which has Arabic roots. A growing body of literature has investigated Islamic law from a technical perspective. However, from the perspective of linguistics and translation studies, little attention has been paid to the lexicon that makes up this specialized discourse. Much of the commentary has so far been prescriptive, with limited empirical evidence. This study aims to bridge this gap by exploring how ‘culturalese’ (i.e., ostensive cultural discourse) travels through language, as evidenced in a self-built monolingual English corpus of over 9 million words, covering diverse genres or registers on Islamic financial and family law matters.

Using a mixed methods design, the study first quantifies the different linguistic strategies used to render sharia-based concepts in English, in order to explore ‘translation’ norms based on linguistic frequency in the corpus. This quantitative analysis employs two models: profile-based correspondence analysis, which considers the probability of lexical variation in expressing a conceptual category, and logistic regression (using MATLAB programming software), which measures the influence of the explanatory variables ‘register’ and ‘legal function’ on the choice between an Arabic loanword and an endogenous English lexeme, i.e., a close English equivalent. The findings are then interpreted qualitatively in light of postcolonial translation agendas, which aim to preserve intangible cultural heritage and promote the representation of minoritized groups.

The research finds that the English-language discourse on Islamic law is characterized by linguistic borrowing and glossing, implying an ideologically driven variety of English that can be labelled as a kind of ‘Islamgish’ (blending ‘Islamic’ and ‘English’) aimed at retaining symbols of linguistic hybridity. The regression analysis confirms the influence of contextual factors (register and textual function) on the use of an Arabic loanword versus English alternatives.

**Keywords:** Corpus linguistics, register variation, textual variation, profile-based correspondence analysis, logistic regression, Islamic law and finance, postcolonial studies

[845] ***Conceptualizing Constructions of Space Particle Xià in Mandarin Chinese: A Corpus Based Study.***

Ansley Chi-Lin Hung (National Taiwan Normal University), Alvin Cheng-Hsien Chen (National Taiwan Normal University) and Jen-I Li (National Taiwan Normal University).

**Abstract.** The postposition *xià* in Mandarin represents the meaning of ‘under.’ In daily discourse, speakers use *xià* with either an optional possessive marker *zhī* (e.g., hé zuò ZHĪ XIÀ ‘under cooperation’) or an optional co-verb *zài* ‘at’ (e.g., ZÀI yuè guāng XIÀ ‘under the moon light’). Sometimes, speakers can use all three together to denote the spatial relation of ‘being under’ (e.g., ZÀI zhǎng kòng ZHĪ XIÀ ‘under control’). These alternative constructions may seem synonymous because *zài* is a preposition that marks a static locational connection with neither overt inflection nor particular spatial semantics (Cai, 2013), and *zhī* merely denotes a generic possessive/associative reading. Since true synonyms are rare (Saeed, 2011: ch3), this study aims to investigate the underlying distinctive semantic features of these near-synonymous variants of space particle constructions (SPCs) associated with *xià* in Taiwanese Mandarin: [NP/VP xià], [NP/VP zhīxià], [zài NP/VP xià], and [zài NP/VP zhīxià]. Our data were obtained from the latest version of the Corpus of Contemporary Taiwanese Mandarin (COCT), which is by far the largest representative Mandarin corpora in Taiwan. Around 360,000 concordance lines in the written corpus (the maximal number a user can access) of the four *xià*-related constructions were retrieved through the standard query search engine provided by the corpus provider, National Academy for Education Research. We adopted a collostructional analysis approach (Stefanowitsch and Gries, 2003), namely, multiple distinctive collexeme analysis (Gries and Stefanowitsch, 2004), to determine the most distinctive landmark (LM) head nouns associated with each type of spatial construction. A chi-square test was then employed as a post-hoc analysis to examine the relationship between the SPC alternative and semantic features of the head nouns.

Our analysis yields the following findings. First, each construction has its strong preference for the co-occurring LMs. [Xià] favors neutral nominal entities (e.g., jiǎo XIÀ ‘under foot); [zhīxià] tends to cooccur with abstract lexemes, most of which are statives such as emotions and feelings; [zài\_xià] often cooccurs with positive nouns (e.g., ZÀI zhī chí XIÀ ‘under the support of’) ; [zài\_zhīxià] prefers negative abstract concepts that are either static or dynamic (e.g., ZÀI bào lì ZHĪ XIÀ ‘under violence’; ZÀI bī pò ZHĪ XIÀ ‘under duress’). This result is inconsistent with Cai’s (2013) observation that *zài* is more abstract in terms of meaning. Instead, our findings suggest that one of the *zài* constructions, [zài\_zhīxià], attracts concrete lexemes. In addition, *zhī* constructions attract static content and *zài* elicits semantic prosody. It is concluded that different constructions of one space particle are not as synonymous as previous studies assumed and elicit specific connotations. This study shows great promise for pondering the function of adpositions as well as circumpositions that are semantically synonymous in both Mandarin and other languages.

**Keywords:** space particle, quantitative corpus linguistics, adposition, near-synonym, collostructional analysis, cognitive semantics

[893] ***“When I first started out”: Metadiscourse in Undergraduate ePortfolios.***

Lydia King (Oxford College of Emory University), Rose Delony (Oxford College of Emory University) and Jack Hardy (Oxford College of Emory University).

**Abstract.** Successful undergraduate writing is often characterized by the ability to argue persuasively (Wingate, 2012), a skill that often involves the writer inserting their own beliefs and perspective into the proposition. The use of interactional metadiscourse, or the way writers present themselves in their writing (Hyland, 2019), has been observed across disciplines (Author, 2014). Author (2014) found that undergraduates were sometimes asked to reflect on tasks but rarely wrote extended reflections of semester- or years-long experiences.

One type of writing that requires that latter kind of reflection and argumentation is the personal ePortfolio, which has recently been implemented as a milestone course at our college. This course asks students to create a website that includes a biographical sketch, artifacts of learning, and a reflective narrative. It serves as a culmination of their first two years of college. The course is led by faculty from across the college, but it does not have a specific disciplinary focus nor assessment.

To gain a better understanding of how these Milestone projects compare to other student writing at our college, we collected a corpus of these projects (n =150) and analyzed it using the Authorial Voice Analyzer (Yoon, 2017). We examined three aspects of metadiscourse: self-mentions, stance, and engagement. While some features show obvious differences (e.g., self-mentions) there were surprising similar amounts of use of other features (e.g., boosters).

We conclude by discussing the task of reflection and offer suggestions for instructors to consider its relationship with narration.

**References**

Gardner, S., Nesi, H., & Biber, D. (2018). Discipline, level, genre: Integrating situational perspectives in a new MD analysis of university student writing. Applied Linguistics.

Hyland, K. (2019). Metadiscourse: Exploring interaction in writing (2 ed.). New York, NY: Bloomsbury.

Wingate, U. (2012). ‘Argument!’ helping students understand what essay writing is about. Journal of English for Academic Purposes, 11(2), 145-154.

Yoon, H.-J. (2017). Textual voice elements and voice strength in efl argumentative writing. Assessing Writing, 32, 72-84.

**Keywords:** undergraduate writing, reflective writing, online language, metadiscourse

[963] ***Construction and Analysis of a Map-Based Corpus for Tracking Linguistic Variation and Demographic Characteristic Identification.***

Theodore Manning (City University of New York Graduate Center), Eugenia Lukin (Duke University), Ross Klein (University of Pittsburgh) and Patrick Juola (Duquesne University).

**Abstract.** Project Map Lemon is a linguistic corpus in its second iteration that was created to obtain a baseline corpus for linguistic variation among English-speaking North Americans. The Map Lemon corpus currently houses upwards of 21,000 words across 185 participants, 10+ linguistic backgrounds, and 40+ US states and Canadian provinces. It presents a unique method for linguistic data collection, as a map-based corpus had not been previously developed for the written medium. Analysis using this corpus has revealed the potential to disambiguate gender and sex, as well as region. Further research is currently being conducted to solidify these results.

Map Lemon data was gathered electronically via participants on Prolific writing responses in a Google form. They were given a small monetary reward for participation. Participants were asked to be as detailed as possible. In Experiment I, participants were asked to guide the fictional Chad LemonLover, using a hand-drawn map, to a lemonade stand using whatever cardinal directions or landmarks they desired. In Experiment II, participants were asked for their recipe for making lemonade. Demographic information was then collected, including birthplace, gender identity, assigned sex, linguistic background, etc.

Results from conducting stylometric analysis using K-Nearest Neighbor and part of speech tagging in the Java Graphical Authorship Attribution Program indicate that transgender respondents write most similarly to their gender identity rather than sex assigned at birth. Additionally, using the same analysis methods, a naturally occurring unknown nationality in our responses was most similar to Canadian writers (when compared to American and Canadian authors from the same corpus). The respondent later confirmed they are Canadian, showing that Map Lemon can be used to disambiguate region.

We feel Map Lemon demonstrates a unique data collection method for documenting linguistic variation on the internet, as well significant results for the field of transgender linguistics.

**Keywords:** authorship attribution, demographic identification, transgender linguistics, computational linguistics, stylometry

[972] ***Dealing with Duplicates in Newspaper Corpora.***

Sharon Glaas (University of Birmingham).

**Abstract.** The availability of digital news article archives facilitates the collection of texts for Corpus Assisted Discourse Studies (CADS). However, an issue that receives little attention in the literature is that of duplicate data. While repositories such as Lexis offer removal of texts from search results due to "similarity", the basis for calculation of similarity is unclear.

This paper addresses the issue of duplicates found when building a corpus of Guardian Newspaper articles that referenced the European Union in the six months leading up to the Brexit referendum. Over 60% of c.10,000 articles initially collected for the corpus had some degree of duplication. The existence of such a large number of duplicates has significant implications for our understanding of media text production Duplication also raises questions as to the consumption of these articles by readers and the extent to which duplication increases exposure to particular ideas and adds to the incremental effects of the representations found.

I outline four distinct types of duplicate found and propose a novel approach to address the challenges of duplicate data in newspaper corpora. I caution against a fully automated approach to deduplication in CADS and instead suggest a semi-automated method for dealing with duplicates that balances the need to address the issue as efficiently as possible but at the same time ensure that appropriate decisions are made with regard to dealing with duplicate data in order to ensure the integrity of subsequent analysis.

**Keywords:** Corpus Building, Duplicate Texts, News Corpora

[1089] ***What can a corpus tell us about workplace creativity?***

Michael Handford (Cardiff University).

**Abstract.** Creativity is regarded as central to a range of workplaces, and one that promises to become increasingly so with the advance of AI (Du Sautoy, 2019). But to what extent can corpus methods help us pinpoint and analyse creativity in workplace interactions?

This talk firstly presents a novel model of creative workplace discourse, comprised of three parameters: linguistic creativity, discursive creativity, and alignment. Features like linguistic metaphors, repetition, jokes and hyperbole are instance of linguistic creativity (Carter, 2016), whereas discursive creativity is less concerned with the formal aspects of language than with the creative uses to which language is put (Jones, 2016). Discursive creativity is therefore the use of discourse to achieve, novel and goal-appropriate social actions, through for instance interdiscursivity, or creative reframing. Alignment is the way interlocutors position themselves towards each other or the task at hand (Hyland, 2012). The three parameters can account for the way workers interactively do creative things, create new identities or practices, or use creative language for potentially rhetorical effect while navigating their relationships.

Employing a CADS methodology, CANBEC (the Cambridge and Nottingham Business English Corpus), a one-million word corpus of authentic business meetings (Handford, 2010) is analysed in terms of creativity. While automated corpus methods, such as keyword lists and Word Sketches (Kilgarriff et al, 2014) can pinpoint some potential features of linguistic creativity, and evidence relevant practices like problem-solving, to examine discursive creativity, alignment and many instances of linguistic creativity with greater precision, it is necessary to manually analyse extended extracts of meetings. Selected extracts will be shown to demonstrate the parameters of the model and the need for manual analysis. The talk will conclude with some reflections on the limitations of current corpus tools for analysing creativity in interactions, and will make some highly tentative suggestions for future tool development.

**Keywords:** creativity, CADS, workplace discourse, critique of corpus tools

[1117] ***Building a corpus of academic writing in EMI contexts: Challenges of collecting data from different educational settings.***

Dana Gablasova (Lancaster University), Luke Harding (Lancaster University), Raffaella Bottini (Lancaster University), Sally Ren (Lancaster University) and Vaclav Brezina (Lancaster University).

**Abstract.** English-medium instruction (EMI) is a major pedagogical trend, reflecting and shaping the status of English as a global language. In EMI contexts, the ability to use academic English – the language through which academic subjects are taught and assessed – is crucial to academic success. However, we know relatively little about the actual use of language in university-level EMI in different countries.

The majority of research on language practices in EMI university contexts is based on self-reported data from questionnaires, surveys and interviews (e.g. Galloway et al., 2020), with only a limited number of studies focusing on the analysis of language produced by the students in EMI contexts (e.g. Pun & Macaro, 2019). The current project, therefore, was designed to provide empirical evidence about the use of academic English in different EMI university contexts and across different subject areas. The project seeks to build a corpus of written university assignments, with data collected from six universities with large EMI provision in China, Italy and Thailand in order to contribute to a better understanding of disciplinary writing at university-level EMI contexts.

The paper focuses on the methodology of corpus creation and discusses the first stage of the project, i.e. data collection. In particular, it addresses the challenges related to collecting a meaningful set of written data and metadata from different countries and from different academic environments. Specifically, the paper focuses on three key issues: a) addressing the subject-related and institution-related differences in written genres (e.g. the typical length of written pieces), b) addressing the institution-related differences in practices involved in producing written assignments (e.g. the preference for individual or pair/group work), and c) collecting metadata about the written assignments given the difference in terminology and practices across the institutions. The paper will discuss the strategies adopted in the international data collection and implications for building a representative and balanced EMI corpus.

**Keywords:** corpus development, English-medium instruction (EMI), data collection, L2 writing

[1139] ***More than words: multimodal multidimensional analysis of popular music in English language.***

Maria Claudia Delfino (Pontifícia Universidade Católica de São Paulo/Fatec Praia Grande).

**Abstract.** Popular music has been the object of several studies in such fields as Musicology, Literature, and Corpus Linguistics, among others. While there is great interest in the study of popular music, the focus of most research has been generally restricted to theoretical concerns, genre or register variation, and a monomodal, small data approach. Unlike these trends in the previous literature, in this paper, I report on a study that relied on a large-scale dataset comprising both the verbal (lyrics) and the musical (voice and instruments) semiotic modes. The major goal of the study was to determine the multimodal characteristics of popular music in English, from a multidimensional (MD) discourse-based perspective (Berber Sardinha & Fitzsimmons-Dooley, in prep.). For the verbal analysis, the basic MD methodology included collecting a 200K-text corpus of lyrics by scraping a popular web repository, automatically annotating the lyrics with semantic information through the USAS tagger, tabulating the frequencies of the semantic categories, entering the counts in a factor analysis, scoring each lyric on each factor, and labeling the factors as dimensions based on their discourse characteristics, through the careful qualitative interpretation of hundreds of lyrics. For the acoustic analysis, the procedures included scraping the Spotify database for the acoustic indexes of each song, entering the indexes as variables in a factor analysis, scoring each music on each factor, and interpreting the resulting factors as acoustic dimensions, based on the qualitative analysis of the musical aesthetics of each song. The multimodal dimensions were obtained through a canonical correlation analysis of the dimension scores for the lyrics and the acoustics, corresponding to the major shared patterns of text and music, namely: Dim.1: Non-mystical, materialistic discourse accompanied by bodily-movement inspiring musicality; Dim.2: Subjective, supernatural discourse accompanied by high energy, intense musicality; and Dim.3: Tragic-romance discourse accompanied by joyful musicality.

**Keywords:** Multidimensional analysis, Music, Big Data, Corpus Linguistics, Multimodality

[1231] ***Xenophobic representations against China on Twitter.***

Cicero Soares da Silva (Pontifical Catholic University of São Paulo), Aline Zamboni Milanez (Pontifical Catholic University of São Paulo), Amanda Chiarelo Boldarine (Pontifical Catholic University of São Paulo) and Eduardo Silva (Goiás State University).

**Abstract.** In this paper, we look at malicious representations of China on Twitter occurring in the context of the COVID-19 pandemic. The pandemic has heightened ongoing prejudice against China, its people and culture partly because the coronavirus originated in that country, and partly because the West has engaged in demoralizing campaigns against China for decades. In order to capture these detrimental representations, we scraped a large corpus of tweets in Brazilian Portuguese containing two highly xenophobic hashtags, namely #pestechinesa (the Chinese pest) and #viruschines (Chinese virus). These were used by right-wing followers in Brazil to discredit China and spread hatred. Unlike North America and Europe, the leading COVID-19 vaccine in Brazil was produced using Chinese know-how, which the right-wing federal administration opposed, hindering its distribution to vaccination centers around the country. The method consisted of an application of Lexical Multidimensional Analysis (Berber Sardinha, 2019; Berber Sardinha & Fitzsimmons-Dooley, in prep; Fitzsimmons-Dooley, 2019, forthcoming; Clarke et al. 2022), a framework for the analysis of discourse in corpora that uses lexical units to detect traces of discourses across the texts. The corpus comprising +100K tweets was tagged for part-of-speech and lemmatized. The lemma counts were entered in a factorial extraction, which yielded factors that were interpreted discoursally as dimensions of variation reflecting xenophobia. The dimensions will be presented and discussed in the paper presentation. In addition to the Lexical Multidimensional Analysis, a machine learning model was trained on a text sample that had been previously labeled for hateful content by hand. The human annotation took place in several rounds and measured using the Kappa coefficient for both intra-rater and inter-rater agreement, reaching at least .7 agreement. The hand-labeled data was used to generate a prediction model using the fasttext classifier, which produced a model achieving more than 90% accuracy in detecting xenophobic content.

**References**

Berber Sardinha, T., & Fitzsimmons-Doolan, S. (in prep.). Lexical Multidimensional Analysis. Cambridge University Press.

Berber Sardinha, T. (2019). Using multi-dimensional analysis to detect representations of national culture. In T. Berber Sardinha & M. Veirano Pinto (Eds.), Multi-dimensional analysis: Research methods and current issues (pp. 231-258). Bloomsbury / Continuum.

Clarke, I., McEnery, T., & Brookes, G. (2021). Multiple Correspondence Analysis, newspaper discourse and subregister: A case study of discourses of Islam in the British press. Register Studies, 3(1), 144-171.

Fitzsimmons-Doolan, S. (2019). Language ideologies of institutional language policy: Exploring variability by language policy register. Language Policy, 18(2), 169-189.

Fitzsimmons-Doolan, S. (forthcoming). 21st century ideological discourses about US migrant education that transcend registers. Corpora, 18(2).

**Keywords:** Xenophobia, Discourse analysis, Lexical Multi-Dimensional Analysis

[1234] ***A corpus pragmatic analysis of politeness in spoken English business communication.***

Isolde van Dorst (Vienna University of Economics and Business).

**Abstract.** British culture is said to be characterised by off-record or negative politeness (Blum-Kulka, 1987) and norms giving prominence to social distance (Bargiela-Chiappini & Kádár, 2011). Work constituting variational pragmatics (Barron & Schneider, 2008) has shown that politeness differs across speakers and contexts. Culpeper and Gillings (2018) found that there are differences in what politeness type is used depending on one’s sociolinguistic background.

In the present study, I investigate whether a nuanced politeness pattern based on linguistic variables can also be found in a specialised (in this case, business) context using the 1-million-word Cambridge and Nottingham Business English Corpus (CANBEC) (Handford, 2010). The corpus contains rich (sociolinguistic) speaker metadata, allowing for various comparisons, for example between genders and professions. Crucially for pragmatic analysis, CANBEC also contains metadata on the context of the interaction (Handford, 2010).

I selected 50 key British formulaic politeness expressions, each allotted to one of three types of politeness (tentativeness, deference or solidarity), and differing levels of formality (e.g., relatively formal goodbye versus relatively informal bye). The selection of expressions was driven by academic literature and/or non-academic literature (e.g., newspaper articles complaining about politeness); my own knowledge of politeness practices; or extrapolations from other expressions (e.g., thank you would lead to a consideration of thanks, ta, cheers). Instances of these 50 expressions were retrieved from CANBEC, and then manually screened to remove non-genuine cases of politeness (e.g., sarcasm).

I applied a linear mixed-effects model to analyse the effect of each social variable on the use of politeness expressions. Clear differences across politeness types and levels of formality emerged. There is also an observable difference in politeness usage based on job level, and between speakers of different ages. There are likely to be multiple and complex reasons for each finding; I will offer some tentative possibilities.

**References**

Bargiela-Chiappini, F., & Kádár, D. (eds.). (2011). Politeness Across Cultures. London: Palgrave Macmillan.

Blum-Kulka, S. 1987. “Indirectness and politeness in requests: Same or different?” Journal of Pragmatics, 11, 131-146.

Culpeper, J. & Gillings, M. (2018). “Politeness variation in England: A North-South divide?” In V. Brezina, R. Love, and K. Aijmer (Eds.), Corpus Approaches to Contemporary British Speech: Sociolinguistic studies of the Spoken BNC2014. New York: Routledge.

Schneider, K. P., & Barron, A. (eds.). (2008). Variational pragmatics: A focus on regional varieties in pluricentric languages (Vol. 178). John Benjamins Publishing.

Handford, M. (2010). The Language of Business Meetings. Cambridge: Cambridge University Press.

**Keywords:** corpus linguistics, politeness, pragmatics, business communication

[1317] ***Variation in learner writing and a comparison with textbook models.***

Marilisa Shimazumi (Cultura Inglesa College at Sao Paulo) and Tony Berber Sardinha (Pontifical Catholic University of Sao Paulo).

**Abstract.** In this paper, we present a novel Multidimensional Analysis (MD) of English learner writing based on the International Corpus of Learner English (ICLE) version 3 and try to determine the extent to which the argumentative learner writing in ICLE corresponds to samples of argumentative writing found in EFL/TESOL textbooks. Our MD analysis was carried out on the complete version 3 of the ICLE corpus, comprising 9,529 texts, 5.7 million words, from 27 different national subcorpora. The corpus was fully tagged for lexicogrammatical features with the Biber Tagger. The feature counts were entered in a factor analysis, revealing two factors, which were interpreted communicatively as dimensions of variation, namely: Dim. 1: Objective, fact-based exposition; and Dim. 2: Subjective, involved exposition. A text typology was developed based on these dimensions, which revealed four distinct text types, corresponding to gradations of the two dimensions. Text type 1 comprised very subjective and somewhat evaluative texts; type 2 included subjective essays not marked for evaluation; type 3 corresponded to fact-based essays with subjective positioning. And type 4 included very objective oriented essays. Overall, the results showed that argumentative writing varies widely around the world, with two very different styles competing, namely a more objective, fact-based style, and a more subjective, opinion-based style. To meet our second goal, we collected a corpus of essays found in textbooks for learners of English, corresponding to levels B2 or upper in the CEFR or Advanced Mid and upper in the ACTFL. This corpus was tagged with the Biber Tagger, enabling a comparison with the learner essays. Comparisons were carried out between each national variety and the textbook samples to determine the extent to which the learners were dimensionally close to or distant from the textbook models. These comparisons will be presented and discussed in the paper.

**Keywords:** Learner writing, Multidimensional Analysis, Argumentative writing

[1402] ***“Fascinating differences”: Problem-oriented language in professional language corpora.***

Rachele De Felice (The Open University).

**Abstract.** This presentation stems from a wider project on how problems are discussed in the workplace (problem-oriented language, henceforth POL), using corpora such as the Clinton Email Corpus (CEC) to uncover authentic communication patterns. One of the project’s aims is to develop pedagogical and training resources for professional communication rooted in real-world language. The many handbooks in this field aimed at the general market usually rely on artificial examples. Arguably, however, artificial examples represent behaviours that are socially desirable and contextually appropriate, which leads to our research questions:

1) Is there overlap between workplace language presented in handbooks and that occurring in corpora?

2) How are socially desirable behaviours instantiated linguistically in corpora?

To investigate this, we compare the CEC to a purpose-built ‘Handbook Corpus’ (HC). The CEC, containing over 35,000 messages exchanged between Hillary Clinton and dozens of correspondents while she was US State Secretary, is a rich source of instances of POL (AnonymisedRef). The HC contains model phrases and dialogues extracted from bestselling books such as The Conflict Resolution Phrasebook ("2000 phrases for… anyone who has to deal with difficult workplace situations"). Frequent phrases and patterns in both corpora are compared to determine areas of similarity.

Initial analysis shows notable divergences. The HC features highly descriptive nouns (e.g. contradiction, problem) and evaluative adjectives (e.g. this is offensive, our differences are fascinating). These are extremely infrequent in the CEC, suggesting interlocutors prefer more neutral or indirect expressions in real-world POL. Conversely, the two datasets share frequent use of modal verbs – especially those indicating obligation (we/they should) and irrealis statements (I/it/we should have), pointing to the assignment of responsibility as a key aspect of POL. Closer qualitative analysis will follow to understand how (or if) the emotions represented by, for instance, evaluative adjectives are conveyed through other means in the CEC.

**Keywords:** professional communication, problem-oriented language, email corpora

[1498] ***An application of Learner Corpus Αnalysis in grammatical comprehension.***

Katerina Florou (National and Kapodistrian University of Athens).

**Abstract.** Two decades after Granger et al. (2002), described learner corpora as “electronic collections of spoken or written texts produced by foreign or second language learners in a variety of language settings” and some years after Ellis (2006) proposed “focusing on form” during any phase of a TBL framework, the present study addresses a research question: how can learner corpora, error annotated or not, provide opportunities to focus on form in an educational environment?

Willis (2016) claims that one of the aims of analyzing the form is to help students explore language and to develop an awareness of different aspects of the foreign language. We can find the same purpose in analyzing error tagged learner corpora or row learner corpora, but in this case we focus on the incorrect forms. According to Ellis (2006) the teacher “should seek to address errors or gaps in the students’ L2 knowledge”, while performing a task. Following this suggestion we compared the written production of two groups of students: the first (G1) was a group that applied the focusing on form phase by observing a particular form of the Italian language, the past participle, in examples from native speakers and in grammar rules and the second (G2) was a group of students that applied the focusing on form phase by observing their own errors and comparing them with corresponding examples of a native speaker corpus. In order to investigate our hypothesis we developed four learner corpora. The first two (one of each group of students) were produced during the task phase (Willis 2016) and the other two at the end of the lesson as a follow up activity. The results of the comparison showed that students’ exposure to their own errors can improve the comprehension of a grammatical element.

**Keywords:** Learner Corpus Analysis, Focus on form, Italian Language as FL

[1537] ***Migrating across times and cultures. Metaphorical images of migration in U.S.A and Italian newspaper discourse between 20th and 21st century.***

Dario Del Fante (Università degli Studi di Ferrara).

**Abstract.** The proposed research work is strongly related to two essential features of human life: migratory movements and metaphors. Migratory movements can be undoubtedly defined as an enduring characteristic of our global history. As history and literature, from the Homeric poems to the Bible, testify, migration is an inherent feature of humanity. Metaphor is more than a linguistic and stylistic phenomenon (Lakoff and Johnson 1980). It is primarily a feature of human language ability (Semino 2008). Considering that metaphors are one of the means through which people organize their experience of life and that newspapers considerably influence people’s perspectives on reality and opinions (Fowler 1991; Marchi 2019), by analyzing the metaphorical representation of migrants in newspapers we might define, to some extent, how people perceive, think of and ultimately experience this group of people. This study investigates the metaphorical representation of migrants through a corpus-assisted newspaper discourse analysis from both a cross-linguistic and historical perspective.

This project consists of a comprehensive corpus-assisted discourse analysis (Taylor & Marchi 2018) of newspaper articles published within two specific time periods, namely 1900-1914 and 2000-2014, to investigate how media discourses around migration have changed and differ between two cultures and across time, specifically focusing on the metaphors used to represent migration as a phenomenon. The metaphors have been extracted from the corpora by adopting a semi-automatic approach based on the analysis of collocations (Steen et al. 2010). Results suggest that migration is conceptualized along similar lines – i.e. as a threat to national security in both Italy and the US in the 2000s, and in the US discourse in the 1900s. By contrast, migration seems to be conceptualized by means of positive metaphors in Italy in the 1900s as a resource for the country. Moreover, it suggests that the perspective of the speaker is a greater determinant of metaphorical representation than historical context and the linguistic or cultural context.

**References**

Fowler, R. (1991). Language in the News: Discourse and Ideology in the Press. London: Routledge.

Lakoff, G., & Johnson, M. (1980). Metaphors we live by. Chicago: University of Chicago Press.

Marchi, A. (2019). SELF-REFLEXIVE JOURNALISM: A corpus study of journalistic culture and community in the Guardian. London: Routledge.

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

Steen, G. J., Dorst, A. G., Herrmann, J. B., Kaal, A., Krennmayr, T., & Pasma, T. (2010). A Method for Linguistic Metaphor Identification: From MIP to MIPVU. Amsterdam: John Benjamins Publishing Company. doi: 10.1075/celcr.14

Taylor, C., & Marchi, A. (Eds.). (2018). Corpus approaches to discourse: A critical review. Milton Park, Abingdon, Oxon ; New York: Routledge.

**Keywords:** corpus-assisted discourse studies, corpus linguistics & metaphor studies, corpus-assisted metaphor analysis, conceptual metaphor theory, migration discourse, historical linguistics

[1549] ***Discourses of police and policing in newspaper coverage of UK climate change protests, 2019-2022.***

Charlotte-Rose Kennedy (Nottingham Trent University) and David Wright (Nottingham Trent University).

**Abstract.** Recent years have seen increased scrutiny of police action at protests and demonstrations. On the one hand, police have been criticised for using excessive force when policing protests (Gilmore et al. 2019). On the other, legislation such as the Police, Crime, Sentencing and Courts Act 2022 (PCSC) has been passed to further increase the range and severity of conditions police can impose on demonstrations (gov.uk 2021).

This paper investigates the extent to which these opposing positions are (re)produced by mainstream UK newspapers, and what implications the coverage of protest policing may have on future public order legislation.

We collected a dataset of 1,994 articles (1,630,120 words) covering the climate change protest groups Extinction Rebellion, Insulate Britain and Just Stop Oil between 2019 (the year the PCSC was introduced) and 2022 (the year the PCSC was implemented). Using corpus-assisted critical discourse analysis, we employed a concordance analysis of the word ‘police’ and identified emergent representational themes. Our findings reveal that while some coverage suggests that the policing of protests is either too harsh or proportionate, the dominant discourse is that the police are ineffectual and that force against protests should be increased. We conclude by arguing that despite the PCSC already increasing police powers, the overarching representation of a ‘light touch’ police force – ‘hamstrung’ by the law and ‘victimised’ by climate change protesters – could have legitimised Parliament’s recent reintroduction of the previously rejected Public Order Bill, legislation that seeks to further limit the UK’s Freedom of Assembly and Expression rights.

**References**

Gilmore, J., Jackson, W. and Monk, H. (2019). ‘That is not facilitating peaceful protest. That is dismantling the protest’: anti-fracking protesters’ experiences of dialogue policing and mass arrest. Policing and Society. 29(1): 36-51.

Gov.uk (2021). Overarching Police, Crime, Sentencing and Courts Act 2022 factsheet. Gov.uk [online]. 9 March 2021. Available: <https://www.gov.uk/government/publications/police-crime-sentencing-and-courts-bill-2021-factsheets>.

**Keywords:** Corpus-assisted critical discourse analysis, climate change, protests, police, policing, public order legislation

[1574] ***Future? Tense!: Exploring the Future in a climate change non-fiction corpus.***

Cinzia Bevitori (University of Bologna) and Jane Helen Johnson (University of Bologna).

**Abstract.** As one of the most controversial issues of our times, climate change has received increased scholarly attention over the last decades in a vast range of genres and from many different theoretical and methodological perspectives. Climate change plays a hegemonic position in imagining social life and the future of the human species (Hulme 2011). Moreover, cross-disciplinary approaches to the representations of the future, such as imaginaries, expectations, speculations, and narratives (e.g. Bryant and Knight 2019, Bazzani 2022) are equally important in terms of how we communicate climate change, and how this will impact citizens’ behaviour (Bevitori 2011, Russo 2019). While a number of scholars have highlighted the crucial role of the media in increasing citizens’ awareness, knowledge, and attitudes towards forms of action with regard to climate change (e.g. Carvahlo 2010, O’Neill and Boykoff 2010), an investigation into non-fiction works in this area has remained largely unexplored (Dahlstrom 2014).

The aim of this paper is thus to fill the gap by investigating a corpus of climate change non-fiction in order to explore discourses of the future (e.g. Neiger 2007, Flottum et al. 2014). As a highly composite genre, including articles, essays and memoirs, climate, and environmental non-fiction can arguably be considered a powerful form of storytelling (Dahlstrom 2014, Nanson 2021), playing a crucial role in science communication and setting policy agendas.

A purpose-built corpus of over two million words consisting of over twenty non-fictional works on climate change and the environment from the 1960s to the present day is used firstly to examine the linguistic and discursive patterns of meaning as regards the future, adopting a Corpus-Assisted Discourse Studies approach (Partington et al. 2013), also incorporating cross-disciplinary perspectives. Secondly, it will provide some discussion about the methodological and interpretative challenges of investigating the future within this particular context.

**References**

Bazzani, G. 2022. Futures in Action: Expectations, Imaginaries and Narratives of the Future, Sociology 2022 (0) 1-16.

Bevitori, C. 2011. “Imagine, If You Will...” Reader Positioning on Climate Change in US Op-Ed Articles”. In G. Di Martino, L. Lombardo, and S. Nuccorini (Eds), Challenges for the 21st Century: Dilemmas, Ambiguities, Directions, Roma: Edizioni Q, pp. 367-76.

Bryant, R. and Knight, D. M. 2019. The Anthropology of the Future. Cambridge: CUP

Carvalho, A., 2010. Media(ted)discourses and climate change. A focus on political subjectivity and (dis)engagement. WIREs Climate Change 1 (2), 172–179

Dahlstrom, M.F. 2014. Using narratives and storytelling to communicate science with nonexpert audiences. Proceedings of the National Academy of Sciences of the United States of America 111 (Suppl. 4): 13614–13620.

Fløttum, K., Müller Gjesdal, A., Gjerstad, O., Koteyko, N. and Salway, A. 2014, Representations of the Future in English Language Blogs on Climate Change, Global Environmental Change, 29, 213-222

Hulme, M. 2011. Reducing the Future to Climate: A Story of Climate Determinism and Reductionism, Osiris, The University of Chicago Press 26:1, 245–266.

Nanson, A. 2021. Storytelling and Ecology: Empathy, Enchantment and Emergence in the Use of Oral Narratives, Bloomsbury Academic, London

Neiger, M. 2007. Media oracles: The cultural significance and political import of news referring to future events, Journalism 2007; 8; 309-321.

O’Neill, S. J.and Boykoff, M. 2010. The role of new media in engaging the public with climate change. In Whitmarsh L., O’Neill S. J., Lorenzoni I. (Eds), Engaging the public with climate change: Communication and behaviour change. London, UK: Earthscan, pp. 233-251.

Partington, A., Duguid. A. and Taylor C. 2013. Patterns and meanings in discourse. theory and practice in corpus-assisted discourse studies. Amsterdam/Philadelphia: John Benjamins.

Russo K. E. 2019. Speculations about the future: Populism and climate change in news discourse, in Populist Discourse. Critical Approaches to Contemporary Politics, E. Hidalgo-Tenorio, M.A.Benítez-Castro, F. De Cesare (Eds), pp.190-206.

**Keywords:** Future, climate change non-fiction, corpus-assisted discourse study

[1588] ***Creating an Open-Access Thesaurus of the Welsh Language Using Natural Language Processing.***

Mahmoud El-Haj (Lancaster University), Nouran Khallaf (Lancaster University), Elin Arfon (Cardiff University), Jonathan Morris (Cardiff University) and Dawn Knight (Cardiff University).

**Abstract.** Despite a wide range of resources available in English and other global languages, the availability of a digital thesaurus of contemporary Welsh and other minoritized languages remains limited. In this paper, we discuss the development of an open-access online thesaurus of the Welsh language that uses natural language processing (NLP) to compare synonyms generated from the National Corpus of Contemporary Welsh (CorCenCC) and other datasets.

The thesaurus is generated using word embeddings, which involves the transformation of words in a corpus to a vector. Words that are similar in meaning or association are closer in the vector space, and embeddings can be used to map the various links between individual lexemes. In this work, word embeddings were generated from CorCenCC data and compared with a gold-standard dataset compiled by native Welsh speakers.

Two experiments were conducted to create the thesaurus resource. In the first experiment, Welsh word embeddings were generated using fastText and Welsh fine-tuned word embeddings, and compared using a selected list of 250 gold-standard words. fastText is a multilingual word embedding for 157 languages, which allows it to generate word embeddings for rare and out-of-vocabulary words because even if words are rare. The Welsh fine-tuned word embeddings were trained on a corpus of 92,963,671 words from 11 sources.

In the second experiment, a gold standard list was created by combining the input of Welsh speakers who evaluated the resource through a crowdsourcing platform. Human evaluators were recruited to refine the output of the two experiments. The combinations of various participants' entries are used to create a gold standard list for the selected 250 words.

The results of these experiments are reported, including any refinements made to the resource based on human evaluation. The resulting open-access thesaurus will be a valuable resource for the Welsh and other minoritized languages.

**Keywords:** Welsh language, thesaurus, word embeddings, natural language processing

[1601] ***Computer-assisted concordance reading.***

Stephanie Evert (Friedrich-Alexander-Universität Erlangen-Nürnberg), Natalie Finlayson (University of Birmingham), Michaela Mahlberg (University of Birmingham) and Alexander Piperski (Friedrich-Alexander-Universität Erlangen-Nürnberg).

**Abstract.** As one of the most fundamental and central techniques of corpus linguistics, concordance analysis supports the identification of recurrent patterns across the occurrences of a search term, phrase or construction (the ‘node’). This is achieved by organising concordance lines according to similarities that become visible through the ‘kwic’ (keyword in context) display format. The key challenge is that the underlying notion of ‘similarity’ is often not clearly defined, and different research questions and applications demand a focus on different aspects of similarity. Typically, choices for

the organisation of concordances are determined by the intuition of experienced analysts, but also strongly driven by the options offered by specific concordance software tools – in particular sorting the right or left context of the node alphabetically. An early approach towards a systematic account of ‘reading concordances’ was proposed by Sinclair (2003), but these ideas have only selectively been taken forward and concordance reading is still not being taught methodically in the corpus linguistics curriculum.

In this talk we want to look at opportunities for enhancing concordance analysis with suitable computational algorithms. Based on examples of existing corpus tools and case studies, we will review current practice in corpus linguistics to arrive at an understanding of how the affordances of current tools work together with qualitative interpretation. We will outline what we see as the fundamental tool-independent principles of ‘selecting’, ‘ranking’, ‘clustering’ and ‘sorting’, and demonstrate what we consider to be useful applications of these principles. Our talk will be illustrated with textual examples from corpora of fiction from both English and German authors. Functionalities we will specifically discuss in this case study build on our previous work on CLiC (Mahlberg et al. 2020) and IMS Corpus Workbench (Evert & Hardie 2011).

**References**

Evert, S. & Hardie, A. (2011). Twenty-first century corpus workbench: Updating a query architecture for the new millennium. In Proceedings of the Corpus Linguistics 2011 Conference, Birmingham, UK. Software available at cwb.sourceforge.io.

Mahlberg, M., Stockwell, P., Wiegand, V. and Lentin, J. (2020) CLiC 2.1. Corpus Linguistics in Context, available at: clic.bham.ac.uk.

Sinclair, J. (2003). Reading Concordances. An Introduction. London: Pearson/Longman.

**Keywords:** concordances, algorithms, methodology, search functionalities, corpus tools

[1624] ***Historical and Contemporary Perspectives of Disinformation.***

William Dance (Lancaster University).

**Abstract.** Disinformation, false information that is shared to deceive others, is a prominent global issue with more than 50 governments worldwide enacting or proposing legislation, task forces and investigations to address its spread (Funke and Flamini, 2019). Despite these actions, little research has investigated the history of disinformation as a linguistic term. To start addressing this gap, this paper traces terms such as ‘disinformation’ and ‘misinformation’ through large historical corpora to address the following two research questions:

1) How far back can these terms be traced?

2) What other synonyms and near-synonyms have been used throughout the centuries?

Historical corpora offer researchers a wealth of qualitative evidence for how linguistic terms are used in context (Mair, 2009) and can present otherwise inaccessible records in the form of machine-readable documents. This research finds that written documents addressing ‘false news’ date back centuries and discovers several competing terms. These synonyms are examined for their frequency over time to see which variants had greater longevity and which were relatively transient.

These findings are compared to contemporary social media datasets of discussions of ‘disinformation’ and ‘misinformation’ to understand how these topics are discussed in the modern era. Using 20-million tokens of Twitter discussions, trends are identified and explored over a ten-year period to identify what dominates discussions of disinformation and how people in ‘real-world’ settings talk about deception online. Combining corpus-based approaches such as keyness and collocation analysis with close qualitative readings of texts, we can better understand how disinformation has changed over time, and what this can teach us about tackling its spread today and in the future.

**References:**

Funke, D and Flamini, D. (2019). A guide to anti-misinformation actions around the world. Poynter. Retrieved from: <https://www.poynter.org/ifcn/anti-misinformation-actions/>.

Mair, P. (2009). Corpora and the study of recent change in language. In A. Lüdeling and M. Kytö (Eds.), Corpus Linguistics, Volume 2 (pp. 1091-1108). De Gruyter: Berlin.

**Keywords:** Disinformation, Deception, Computer-mediated communication, Twitter

[1633] ***Language complexity of children’s fiction: a contrastive study of the impact of the reader, writer and language.***

Marketa Mala (Univerzita Karlova, Praha).

**Abstract.** 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. Complexity and readability tests (Bulté & Housen 2012; Crossley, Allen & McNamara 2011) can reveal quantitative differences between fiction for children and for adults, but they may obscure the impact of authors’ linguistic idiosyncrasy, and can hardly be used to compare typologically distinct languages (Barbieri Durão & Kloeppel 2018). Taking such tests merely as its starting point, the paper explores linguistic complexity at the local level (Sinnemäki 2014), i.e. the structure complexity of individual linguistic features, “arising from the number of linguistic elements and their interrelationships” (Pallotti 2014: 1). It focuses on supplementive participial clauses, whose processing complexity is due to their “implicit and somewhat ill-defined relationship with the main clause” (Biber et al. 1999: 782-3), semantic indeterminacy (Hasselgård 2010), and unexpressed subject, particularly in sentence-initial clauses. The analysis draws on a corpus compiled to allow comparison along the dimensions of the intended reader, the author and language. The corpus comprises English fiction books written for children (303,000 words) and for adults (630,000 words) by two authors (J. K. Rowling, R. Dahl), and translated into Czech. The results show individual variation in local complexity: in books for both types of reader, Rowling uses a significantly higher number of supplementive clauses per sentence than Dahl. The reader, however, constitutes a factor leading to the reduction of length and internal complexity (number of dependent clauses) of supplementive clauses in her children’s novels. The regard to the reader appears to influence the Czech translations too. The tendency towards reducing structure complexity in children’s (compared to adult’s) fiction is reflected in the preference for more explicit translation counterparts (initial subjects, clause connectors, finite verbs).

**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.

Hasselgård, H. 2010. Adjunct Adverbials in English. Cambridge: Cambridge University Press.

Pallotti G. 2014. A simple view of linguistic complexity. Second Language Research. 1-18.

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

Sinnemäki, K. 2014. Complexity trade-offs: a case study. In Newmeyer, F.J. & L.B. Preston (eds) Measuring Grammatical Complexity. Oxford: Oxford University Press.179-201.

**Keywords:** language complexity, children's fiction, translation, supplementive clauses, English/Czech

[1649] ***Shame on you! A corpus-based analysis of green marketing discourse in mission statements of certified sustainable fashion firms.***

Yingnian Tao (Lancaster University) and Mark Ryan (Lancaster University).

**Abstract.** The fashion industry, with its long and complex supply chains that aim to meet the fluctuating demands of consumers, is frequently criticized for its negative environmental and social impacts. This has led to pressure from consumers and governments for the fashion industry to fix these problems, and fashion firms appear to have responded to this challenge, with frequent

mentions of their environmental and social goals in their marketing materials. However, not everyone is buying it; consumers have accused fashion firms of greenwashing, for example, by promoting misleading environmental claims about products or services (Delmas & Burbano, 2011; Seele & Gatti, 2017) or only focusing on environmental actions with marketable outcomes (Wu, Zhang, & Xie, 2020). Greenwashing has been shown to have negative effects on consumers’ perceptions of brands (Akturan, 2018). However, consumers have difficulties determining the credibility of organizations’ green claims (Parguel, Benoît-Moreau, & Larceneux, 2011); this necessitates a better understanding of what genuine sustainability claims look like.

With this research gap in mind, this study aims to uncover how fashion firms communicate sustainability initiatives and goals. Specifically, this study will explore three research questions: 1) How do sustainable fashion firms present their sustainability goals and initiatives on their websites? 2) How do unsustainable fashion firms talk about sustainability on their websites? 3) Based on the previous two questions, what information may indicate that firms may be engaged in greenwashing?

We will use corpus-based discourse analysis method to answer these questions, including a macro-level corpus analysis and a micro-level individual text analysis (cf. Bednarek, 2009). The fashion firms we choose are based on their sustainability ratings provided by Good On You, a sustainable and ethical fashion brand ratings website. Good On You rates fashion brands on a scale from 1 (least sustainable) to 5 (most sustainable) by measuring their impact on people, planet, and animals. We considered firms rated 4 and 5 as genuinely sustainable firms, and firms rated 1 and 2 as unsustainable firms. We gathered URLs of any of these companies' webpages that discussed sustainability. We uploaded these URLs to Sketch Engine (a corpus manager and corpus analysis tool) to scrape web texts and built a corpus for further analysis. The corpus consists of two sub-corpora: one for sustainable firms (193,313 tokens) and the other for unsustainable firms1.

A macro-level analysis will involve a few corpus queries, including word frequency lists, concordance lines, and word sketches, to uncover the terms that fashion firms use to advertise their sustainability. The top 100 most frequent words will be categorised according to their semantic similarities (cf. Baker, Brookes, Atanasova, & Flint, 2020; Harvey, 2012). The micro-level analysis will be answered by a thematic analysis of webpage texts of randomly selected 50 good rating companies and 50 bad rating companies. We will code the webpage texts according to two themes: the actions that firms advertise, such as, using organic cotton, and increasing workers’ payment, and the time orientation that action suggests, including past orientation and future orientation (cf. Silva, 2021). A comparative analysis of the results of the macro and micro level analysis will uncover the differences or similarities in how fashion firms communicate sustainability to consumers via their websites. This project’s findings will provide insights into what reputable sustainability claims by fashion brands look like, and will therefore provide consumers with a tool to better evaluate the green claims they encounter in their lives.

**Keywords:** Green marketing, Sustainability, Corpus linguistics, Fashion firms, Greenwashing, Good On You

[1659] ***‘No, I can't blame that on dyslexia or anything else’—a corpus-based discourse study of the representations of dyslexia in the UK press between 2017 and 2021.***

Suhua Tang (Animal Think Tank).

**Abstract.** Dyslexia, which affects 10% of the population (NHS, 2022), is reported to be the most common learning disability in the UK (Handler et al., 2011). Its invisibility along with little public knowledge of the condition means many people access information about dyslexia through the media (Devotta et al., 2013). However, the media representations of dyslexia remain under-explored. To expand the current scholarship on media representations of dyslexia (Kirby, 2019; Simblett, 2021), this study conducts a corpus-based discourse analysis of media representations of dyslexia in the UK press between 2017 and 2021. Specifically, collocates of ‘dyslexia’ and adjectival ‘dyslexic’ are generated with the aid of Sketch Engine, and on its basis concordance analyses are examined to uncover the discourses surrounding dyslexia in nine UK national newspapers, which are subcategorised into the 208,683-word-broadsheet and 127,105-word-tabloid corpora. This

research has identified four prominent discourses across the two corpora, namely, Pathologisation discourse, Survivor discourse, Neurodiversity discourse and Suspicion discourse. The former three discourses have parallels in previous research on media representations of dyslexia (Kirby, 2019; Simblett, 2021) and Suspicion discourse is a new finding. Discussed against Kirby's (2019) findings on media representations of dyslexia, dyslexia as a joke has not been found as a salient discourse. Additionally, the key discourse of cures and treatments of dyslexia in Simblett's (2021) study does not emerge as salient in this study. This change is observed in the media framing of other neurodiverse conditions, such as autism, where the discussion of medical solutions has declined (McKeever, 2012).

**Keywords:** corpus linguistics, discourse analysis, dyslexia, pathologization, survivor, neurodiversity, suspicion

[1736] ***Metaphors, NLP and qualitative work. A metaphor annotated corpus about migration and coronavirus as a model for metaphor detection and classification.***

Dario Del Fante (Università degli Studi di Ferrara) and Salvador Ros Muñoz (Universidad Nacional de Educación a Distancia).

**Abstract.** Since the meaning of a figurative expression is very context-dependent, metaphorical language is a challenging topic for computational modelling [5]. Metaphors are pervasively used in every discursive context [7;11], from religion to medical communication, since they connect the domain of familiar and concrete experiences with the domain of abstract and complex concepts and help people to understand the latter through the former [2]. Thus, metaphors represent a fundamental linguistic resource through which we understand the fabric of reality. Their identification process has been extensively considered in the literature as problematic [12] since metaphors might be interpreted differently by two speakers. Although a qualitative approach might be the safest methodology, it is time-consuming and would ideally require more than two human coders to be effectively reliable. Corpus Linguistics [2;4] has offered some solutions based on the focus on the context through collocations, but the processes showcased still relied mainly on qualitative techniques. Despite many issues which make their automatic identification complex [14], many scholars have recently developed different automatic identification techniques exploiting the last developments offered by AI [1], [10], [15]. Many NLP works have explored the idea of constructing a numerical representation of documents retaining some of their semantic features allowing a computer to work with them as a human does through transformer-based models [3], [9], [10]. Therefore, this work aims to build a model for the automatic detection and classification of metaphors. Then, a multiclass text classification task with three labels - none, covid, and migration – has been performed. Therefore, we built a bilingual Italian-English metaphor annotated corpus - 10.000 words and 345 sentences containing COVID-19 and Migration metaphors. Given a sentence, the model is expected to predict the presence/absence of a metaphor and determine its kind of it. For our experiments, we have used hugginface’s autotrain [1] for fine-tuning several BERT-based multilingual cased language models obtaining the best model accuracy result of 0.81. These results encourage us to further train the model, compare it with an already-existent model used for metaphor detection, and extend the experiment to more metaphorical mappings and more languages.

**References**

1. Bizzoni, Y., & Ghanimifard, M. (2018). Bigrams and BiLSTMs Two Neural Networks for Sequential Metaphor Detection. Proceedings of the Workshop on Figurative Language Processing, 91–101.

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

3. Devlin, J., et al.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

4. Deignan, A. (2005). Metaphor and corpus linguistics. Amsterdam ; Philadelphia: J. Benjamins Pub.

5. Khalid Alnajjar, Mika Hämäläinen and Shuo Zhang (2022) Ring That Bell: A Corpus and Method for Multimodal Metaphor Detection in Videos. In the Proceedings of the Third Workshop on Figurative Language Processing.

6. Mikolov, T., et al. (2013). Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems, 30, pp.3111–3119.

7. Lakoff, G., & Johnson, M. (1980). Metaphors we live by. Chicago: University of Chicago Press.

8. Nacey, S., et al.(2019). Metaphor identification in multiple languages: MIPVU around the world. Amsterdam: John Benjamins Publishing Company.

9. Navarro-Colorado, B. (2015). A fully unsupervised Topic Modelling approach to metaphor identification. In: Actas del XXXI Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural.

10. Rai, S., et al. (2016). Supervised metaphor detection using conditional random fields. In Proceedings of the Fourth Workshop on Metaphor in NLP. pp. 18–27.

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

12. Steen, G. J., Dorst, A. G., Herrmann, J. B., Kaal, A., Krennmayr, T., & Pasma, T. (2010). A Method for Linguistic Metaphor Identification: From MIP to MIPVU. Amsterdam: John Benjamins Publishing Company. doi: 10.1075/celcr.14

13. Tong, X., et al. (2021). Recent advances in neural metaphor processing: A linguistic, cognitive and social perspective. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 4673–4686.

14. Veale, T., et al. (2016). Metaphor: A Computational Perspective. Cham: Springer International Publishing.

15. Wilks, Y., et al. (2013). Automatic Metaphor Detection using Large-Scale Lexical Resources and Conventional Metaphor Extraction. Proceedings of the First Workshop on Metaphor in NLP, pp. 36–44.

This work has been developed within the European Project H2020 - Computational Literary Studies Infrastructure - Grant agreement ID: 101004984.

[1] <https://huggingface.co/autotrain>

**Keywords:** metaphor detection, metaphor classification, automatic metaphor identification process, corpus compilation, NLP, Machine Learning

[1737] ***A Crowdsourcing Annotated Tourism Review Corpus of Emotion.***

Mansour Almansour (Lancaster University), Paul Rayson (Lancaster University) and Scott Piao (Lancaster University).

**Abstract.** We present a tourism review corpus, named TORCE (TOurism Review Corpus of Emotion), which was collected from a major online tourism website and annotated with Plutchik’s eight emotion categories (Plutchik, 1980) via crowdsourcing.

Our survey shows there are very few such corpora with emotion annotation, and our corpus provides a valuable language resource for studying emotion expressions in tourism context.

The raw corpus data was collected using web crawling via the TripAdvisor tourism website (https://www.tripadvisor.co.uk), with their approval. From the raw data, we first selected reviews (some contain multiple sentences) related to attractions in the UK, from which we further selected those reviews that contain at least one emotion word. We also removed excessively long and short reviews, obtaining 2,500 sentences for the crowdsourcing annotation.

In regard to the crowdsourcing annotation, we paid MTurk (<https://www.mturk.com>) workers to classify each review into one or two most likely emotion category/ies from Plutchik’s eight emotion classes: “anger”, “anticipation”, “disgust”, “fear”, “joy”, “sadness”, “surprise”, and “trust”. We assigned the task only to the workers with proven high reputation and expertise. For each review, five workers were asked to annotate it with up to two most likely emotion categories. Inevitably, the crowdsourced annotation contained disagreements, such as incompatible emotion annotations from different workers, which we have analysed in order to reflect on how distinctive the emotion classes are and what information we can potentially learn where people disagree on a classification. To produce a final agreed annotation, the crowdsourced annotation was pruned by applying majority voting approach, i.e. only retaining those emotion classes for each review chosen by the majority of annotators.

We also analysed the inter annotator agreement (IAA), which shows a promising annotation consistency, obtaining a Plutchik Emotion Agreement (PEA) (Desai et al., 2020) score of 0.86. For example, in our corpus, 23% reviews received full agreement on emotion category from 5 workers, while 54% received the agreement from 3 workers. In terms of distribution of emotion categories, “Joy” is the dominant class that covers about 31% of the reviews. In contrast, only 2% of the reviews fall under the “Fear” category. This finding implies that tourism reviews tend to be optimistic and positive. Moreover, our experiment shows that emotion annotation of travel reviews involves subjective judgement and therefore it is a highly challenging task.

**References**

Shrey Desai, Cornelia Caragea, and Junyi Jessy Li (2020). Detecting perceived emotions in hurricane disasters. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5290–5305, Online. Association for Computational Linguistics.

Robert Plutchik (1980). A general psychoevolutionary theory of emotion. In Robert Plutchik and Henry Kellerman (eds), Theories of Emotion, pages 3–33. Academic Press.

**Keywords:** Emotion Annotation, Crowdsourcing, Corpus annotation, Tourism, Natural Language Processing

[1747] ***Which English do we need to teach to pilots? A genre analysis through corpora.***

Malila Prado (BNU-HKBU United International College).

**Abstract.** In aviation English teaching, there is reported resistance to the employment of real communications, in the claim that these are poor English models as they contain ‘noises’ such as disfluencies (false starts, filled pauses) and mistakes, features that find support in written grammar. Due to the high-stakes nature of proficiency exams in the aviation English area, the focus often lies on exam-oriented practices with a restricted scope of language that is inspired by idealized speakers, thus showing no correspondence to real life. This understanding may be attributed to the lack of research on the intersection of ESP, spoken language and World Englishes (see Bolton & Jenks, 2022). In an attempt to identify the characteristics of the English used by pilots, this paper refers to a spoken corpus of radio exchanges between pilots and air traffic controllers in abnormal situations that has revealed a high frequency of pragmatic items such as politeness markers, mitigation (hedging and boosting), and honorifics, indicating rapport engineered through the need to share responsibility over the ongoing problem (Prado, 2021). This corpus is complemented by another, which consists of work-related genres, such as checklists, operating manuals, forums, reports, among other documents (see Friginal, Mathews & Roberts, 2020). Through Corpus Linguistics and Critical Genre Analysis (Bhatia, 2017), this paper investigates what constitutes the dynamic and mobile work setting of pilots, as well as the English

employed in such settings. It also analyzes the extent to which the content of the written material that pilots routinely tackle influences the language that they use when communicating with air traffic controllers. It concludes with pedagogical applications of the technical language present in documents and radio communications as models of real language use to design aviation English lesson plans, decentralizing the view from idealized written grammar.

**Keywords:** aviation English, Corpus Linguistics, ESP, World Englishes, pedagogical applications

[1934] ***Exploring Sinophobic discourses during the COVID-19 pandemic: A corpus-assisted comparative discourse study of Italian, American and Australian newspapers.***

Ilaria Iori (University of Modena and Reggio Emilia).

**Abstract.** Recent investigations on social media uncovered the presence of hateful messages directed at Chinese people during the COVID-19 pandemic (Costello et al., 2021; DeCook & Yoon, 2021; Lee, 2021). Furthermore, corpus-assisted critical discourse studies on Sinophobia in news media (Kania, 2022; Sedláková, 2021) confirmed Van Dijk’s theories (Van Dijk, 1991, 1999) on the reproduction of racism and showed that also Sinophobic messages are often delivered through the discursive construction of a positive Us and a negative Them. This paper aims at contributing to the field of Sinophobic discourses by comparatively exploring the representation of China in Italian, American and Australian newspapers.

From a methodological perspective, the analysis takes a corpus-assisted critical discourse approach (Baker et al., 2013). A newly built corpus (4,673,473 tokens) was used to support the analysis. More specifically, the corpus comprised articles on China published in the first six months of 2020 in the most read Italian, American and Australian broadsheets. Drawing on the previously mentioned studies, the analysis uses #Lancsbox 6 (Brezina et al., 2020) to investigate the collocational profiles of the words China/Chinese and Cina/Cinese and the concordances of personal pronouns and possessives associated with China.

Preliminary results show that China has been frequently represented as a threatening Other in the three corpora. However, the Italian newspapers also associated words with a strong positive connotation, specifically in February and March 2020 (e.g., i nostri amici in Asia/our friends in Asia). This contrasting attitude might suggest that the spread of COVID-19 in the countries under inquiry influenced the lexical choices made by the journalists. The data also indicate that American newspapers take a more critical stance when representing their own institutions, whereas, in the Italian and Australian newspapers, a positive self-representation seems to prevail and reinforces the strong opposition between China and Western countries.

**References:**

Baker, P., Gabrielatos, C., & McEnery, T. (2013). Discourse analysis and media attitudes: The representation of Islam in the British press. Cambridge University Press.

Brezina, V., Weill-Tessier, P., & McEnery, T. (2020). #Lancsbox v.5.x. <http://corpora.lancs.ac.uk/lancsbox>.

Costello, M., Cheng, L., Luo, F., Hu, H., Liao, S., Vishwamitra, N., Li, M., & Okpala, E. (2021). COVID-19: A Pandemic of Anti-Asian Cyberhate. Journal of Hate Studies, 17(1), 108. <https://doi.org/10.33972/jhs.198>

DeCook, J. R., & Yoon, M. H. (2021). Kung Flu and Roof Koreans: Asian/Americans as the Hated Other and Proxies of Hating in the White Imaginary. Journal of Hate Studies, 17(1), 119. <https://doi.org/10.33972/jhs.199>

Kania, U. (2022). “Snake flu,” “killer bug,” and “Chinese virus”: A corpus-assisted critical discourse analysis of lexical choices in early UK press coverage of the COVID-19 pandemic. Frontiers in Artificial Intelligence, 5, 970972. <https://doi.org/10.3389/frai.2022.970972>

Lee, C. (2021). #HateIsAVirus: Talking about COVID-19 ‘Hate’. In Rodney H. Jones (Ed.), Viral Discourse. Cambridge University Press.

Sedláková, R. (2021). Anti-Chinese sentiment in the Czech public service media during the COVID-19 pandemic. Lodz Papers in Pragmatics, 17(1–2), 65–86. <https://doi.org/10.1515/lpp-2021-0004>

Van Dijk, T. A. (1991). Racism and the press. Routledge.

Van Dijk, T. A. (1999). Discourse and Racism. Discourse & Society, 10(2), 147–148. <https://doi.org/10.1177/0957926599010002001>

**Keywords:** Corpus-Assisted Discourse Analysis, Critical Discourse Analysis, Sinophobia, Racist Discourses

[2095] ***“Get the woke intern away from the twitter account stat!!!” Exploring reactions to Ryanair’s digital communication.***

Ursula Lutzky (Vienna University of Economics and Business) and Andrew Kehoe (Birmingham City University).

**Abstract.** Over the last decade, Twitter has established itself as a key platform for external business communication, with the majority of airlines using it to communicate with their stakeholders (Lutzky 2021). While its recent change of ownership may lead to changes in the use and function of Twitter, this paper focuses mainly on the period prior to its takeover by Elon Musk. We use corpus methods to study the communication style of the Irish airline Ryanair, which changed drastically during the first six months of the COVID-19 pandemic. While Ryanair has had a reputation as an edgy airline for some time, its approach to digital communication shifted towards the end of 2020 when its tweets increasingly adopted a stance on socio-political issues. In this paper, we explore a corpus of 35,000 tweets posted by (24%) and directed at (76%) Ryanair between November 2020 and January 2021. Using a combination of corpus methods, including keywords and collocation, we explore the main topics Ryanair engaged with and stakeholders’ reactions to their tweets. Corporate engagement with socio-political issues may lead to the polarisation and politicisation of organisations (van de Meer and Jonkman 2021), and in this paper we aim to uncover stakeholders’ attitudes to Ryanair’s highly polarising approach in its external communication, showing how corpus linguistics can be used to gain insights into the effectiveness of corporate social media strategies.

**References**

Lutzky, U. (2021). The Discourse of Customer Service Tweets. London: Bloomsbury

Van der Meer, T. G. L. A. and J. G. F. Jonkman (2021), ‘Politicization of corporations and their environment: Corporations’ social license to operate in a polarized and mediatized society’, Public Relations Review, 47 (1). <https://doi.org/10.1016/j.pubrev.2020.101988>.

**Keywords:** business communication, airlines, Ryanair, polarisation, social media, Twitter

[2142] ***Bringing corpus use into classroom: The case of a Secondary School Teacher in an EFL context.***

Ece Genç-Yöntem (Yeditepe University) and Evrim Eveyik-Aydın (Yeditepe University).

**Abstract.** Despite the rapid developments in the field of corpus linguistics and Data Driven-Learning (DDL) for the last three decades, corpus use with young learners is still a rare practice in many EFL contexts including Turkey. This is mainly because DDL is not part of teachers’ professional knowledge yet, as it is not covered thoroughly in the initial teacher education programs. This case study aims to explore an EFL teacher’s experience with DDL within the context of a secondary public school. After receiving a four-week training on DDL designed specifically for the purpose of the study as part of an MA course she is enrolled in the English Language Teaching Program, the participating teacher developed materials using corpus tools and concordancers to teach reading, vocabulary and grammar to 8th graders. Data came from the reflective logs kept after each training session, materials developed using corpus tools and concordancers and implemented in a real classroom setting, and the interviews made before and after the training and the implementation of materials. By highlighting the insights into the teacher’s hands-on experience with DDL, the study revealed not only the affordances but also the limitations of the corpus-based DDL with secondary school learners. The implications involved invaluable suggestions to increase EFL teachers’ corpus literacy and corpus based language pedagogy both in pre-service and in-service teacher education. The study also revealed a need to compile a corpus of age- and proficiency-level appropriate texts to be used more effectively with secondary school EFL learners.

**Keywords:** Data-driven learning, corpus-based materials development, EFL learners, secondary school learners

[2173] ***Revealing registerial and ideological shifts in translation via word frequencies and keyword analyses: a case study of Wild Swans and Mao’s Last Dancer.***

Long Li (The Chinese University of Hong Kong).

**Abstract.** Corpus linguistics has been shown to be highly useful in contrastive linguistics (cf. Xiao and McEnery 2006) and translation studies (cf. Baker, 1993). A translation can be compared to corpora of non-translated text to isolate a translator’s idiosyncratic selections from 1) typical features of the target language/register (Steiner, 2002), 2) translationese, or potentially recurring features of the translated language (cf. Laviosa, 2002; Kruger, 2012). This paper presents the findings of a corpus-based investigation of two politically volatile English autobiographies by Chinese migrant writers, Wild Swans (Chang, 1991) and Mao’s Last Dancer (Li, 2003), and their translations “back” to Chinese (Chang 1992; Li 2009). Specifically, it will calculate word frequencies and keywords (Baker 2010; 2017) through a bilingual comparison of English and Chinese. Keywords in the two English source texts (STs) will be compared not only against each other but also to genre-compatible reference corpus – Category G (Belles-lettres, biographies, essays) of the FLOB Corpus of British English (Mair 1999). Likewise, the two Chinese target texts (TTs) will be compared against each other and to a reference corpus – Category G of the LCMC Corpus of Mandarin Chinese (McEnery and Xiao 2004). Lastly, word frequencies and keywords will be compared between each pair of ST and TT to reveal potential patterns of translation shifts. A synthesis of these quantitative results will provide strong evidence that sheds light on similarities and differences in the key representations, motifs, and narratives of two exceedingly successful English autobiographies by Chinese migrant writers and any important shifts and deviations in the Chinese translations. Synthesising these new lexical analyses with the author’s previous corpus-based grammar analyses of modality (Li 2017) and degree of intensity (Li 2018), the paper will elucidate ideologies around the production and reception of such English works and their Chinese translations.

**Primary data**

Chang, J. (1991/2003). Wild Swans – Three Daughters of China. New York: Touchstone.

Chang, J. (1992/2011). 鴻：三代中國女人的故事 [orig. Wild Swans–Three Daughters of China]. Translated by Pu Zhang. Taipei: Heliopolis/Clio Culture.

Li, C. (2003/2009). Mao’s Last Dancer. Melbourne: Penguin Group (Australia).

Li, C. (2009). 毛澤東時代的最後舞者 [orig. Mao’s Last Dancer]. Translated by Xiaoyu Wang. Taipei: China Times.

The Freiburg-LOB Corpus (‘F-LOB’) (original version) compiled by Christian Mair, Albert-Ludwigs-Universität Freiburg.

McEnery, T., and Xiao R. (2004). The Lancaster corpus of Mandarin Chinese (LCMC) (original version). Lancaster University.

**References**

Baker, M. (1993). Corpus Linguistics and Translation Studies: Implications and Applications. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), Text and Technology: In Honour of John Sinclair (pp. 233-250). Philadelphia/Amsterdam: John Benjamins.

Baker, P. (2010). Corpus methods in linguistics. Research methods in linguistics, 93-113.

Baker, P. (2017). American and British English: Divided by a common language?. Cambridge University Press.

Laviosa, S. (2002). Corpus-based Translation Studies: Theory, Findings, Applications. Amsterdam & New York: Rodopi.

Kruger, H. (2012). A corpus-based study of the mediation effect in translated and edited language. Target, 24(2), 355-388.

Li, L. (2017). An examination of ideology in translation via modality: Wild Swans and Mao’s Last Dancer. Journal of World Languages, 4(2), 118-144.

Li, L, and Wu, C. (2019). Degree of intensity in English-Chinese translation: a corpus-based approach, Functional Linguistics (6)3.

Steiner, E. (2002). Grammatical metaphor in translation. Some methods for corpus-based investigations. In H. Hasselgård, S. Johansson, B. Behrens, & C. Fabricius- Hansen (Eds.), Information structure in a crosslinguistic perspective (pp. 213– 228). Amsterdam: Rodopi.

Xiao, R., & McEnery, T. (2006). Collocation, semantic prosody, and near synonymy: A cross-linguistic perspective. Applied linguistics, 27(1), 103-129.

**Keywords:** contrastive linguistics, translation, ideology, autobiographies, Chinese, keyword, word frequency

[2182] ***On register-specificity of personification in Hungarian – a corpus-driven study.***

Gábor Simon (Eötvös Loránd University Budapest).

**Abstract.** Although personification is a well-known phenomenon in stylistics (see e.g., Melion‒Ramakers (eds.) 2016), its linguistic organization is yet to be investigated (see Dorst 2011), especially concerning the Hungarian language. Based on previous research on English personification (Dorst‒Mulder‒Steen 2011), the presentation explores whether register has any effect on the grammatical and semantic structure of personification in online Hungarian texts. Within the methodological framework of the PerSECorp project (a specific research corpus built for analysing personification, see Simon 2022), online Hungarian car reviews (10486 tokens in total) and articles on international politics (7012 tokens altogether) are compared, after a semi-automatic pre-processing with the e-magyar Digital Language Processing System (Indig et al. 2019) and the manual annotation of personifying expression with the INCEpTION semantic annotation platform (Klie et al. 2018). The annotation extends not only to the morphosyntactic structure of the expressions but also their role in personifying meaning generation, their idiomaticity (on the basis of the collocational behaviour of the components) and the semantic relations between the components of the multi-word expressions.

After processing 75% of the whole sample, the results demonstrate differences between the two registers under investigation both in the frequency and organization of personifying expressions. Car reviews encompass more personifications with a richer variety of semantic structures (including larger verb + argument constructions and a higher number of idiomatic expressions). Whereas in articles on international politics, metonymic personifications dominate the sample, and the personifying expressions are more limited regarding the number of components. The study points out that personification is a complex and diverse phenomenon even in journalistic texts, providing a solid base for further analysis of literary and scientific discourse.

**References**

Dorst, Aletta G. 2011. Personification in discourse: Linguistic forms, conceptual structures and communicative functions. Language and Literature 20 [2]: 113 – 135.

Dorst, Aletta G. – Mulder, Gerben – Steen, Gerard J. 2011. Recognition of personification in fiction by non-expert readers. Metaphor and the Social World 1 [2]: 174 – 201.

Indig, Balázs - Sass, Bálint - Simon, Eszter - Mittelholcz, Iván - Vadász, Noémi - Makrai, Márton 2019. One format to rule them all – The emtsv pipeline for Hungarian. In: Proceedings of the 13th Linguistic Annotation Workshop. Association for Computational Linguistics, 2019, 155-165.

Klie, Jan-Christoph ‒ Bugert, Michael ‒ Boullosa, Beto ‒ de Castilho, Richard Eckart ‒ Gurevych, Iryna 2018. The INCEpTION Platform: Machine-Based and Knowledge-Oriented Interactive Annotation. In: Proceedings of System Demonstrations of the 27th International Conference on Computational Linguistics (COLING 2018), Santa Fe, New Mexico, USA, 5‒9.

Melion, Walter S. ‒ Ramakers, Bart (eds.) 2016. Personification. Embodying Meaning and Emotion. Leiden, Boston: Brill.

Simon, Gábor 2022. Identification and Analysis of Personification in Hungarian: The PerSECorp project. In: Calzolari, Nicoletta‒Béchet, Frédéric‒Blache, Philippe‒ Choukri, Khalid‒ Cieri, Chritopher‒ Declerk, Thierry‒ Goggi, Sara‒ Isahara, Hitoshi‒ Maegaard, Bente‒ Mariani, Joseph‒ Mazo, Hélene‒ Odijk, Jan‒ Piperidis, Stelios (eds.): Proceedings of the 13th Language Resources and Evaluation Conference. Paris: European Language Resources Association (ELRA), 2730‒2738.

**Keywords:** personification, annotation, register, online texts

[2259] ***Building a taxonomy of pattern-based constructions.***

Susan Hunston (University of Birmingham).

**Abstract.** Units of form and meaning that incorporate both lexis and grammar have been proposed from several perspectives, including among others Construction Grammar and Pattern Grammar. The compatibility of these two approaches is by now well known and is ripe for exploitation. It has been proposed that the extensive corpus-based research behind grammar patterns can be used to obtain a more systematic set of constructions than currently exists. This paper describes the results of a project to derive a comprehensive database of verb-based mid-level constructions in English. The data for the project comprises the 100 verb complementation patterns identified from corpus-based lexicography, together with all the verbs observed to occur in those patterns. From this data, close to 1,000 constructions are identified. As well as introducing the database and its contents, the paper covers three aspects of the project. First is the finding that system networks, similar to those found in Systemic-Functional Grammar, are the optimum way to codify the relationship between constructions of different levels of specificity. This is illustrated with an example of a clausal pattern (‘V to-inf’) and a preposition-based one (‘V n against n’), and their associated constructions. Second is a discussion of the nomenclature to be used in describing the construction elements and the extent to which this can draw on approaches such as the Pattern Dictionary of English Verbs or FrameNet. Again this is illustrated with a clausal pattern (‘V n that’) and a preposition-based one (‘V on n’), and the constructions derived from them. Finally the paper tackles the major problem in this work: how to identify and classify the constructions associated with the straightforward transitive and intransitive patterns (‘V n’ and ‘V’); and discusses whether the same construction-based approach can account for this central aspect of English clause structure.

**Keywords:** Construction Grammar, Pattern Grammar, Verbs in English, Corpus-based lexicography, Form and meaning

[2399] ***Is multi-dimensional analysis replicable? Variables, parameters and transparency.***

Elen Le Foll (UCLouvain) and Vaclav Brezina (Lancaster University).

**Abstract.** The study of registers represents a major contribution to our understanding of functional variation in language. Multi-dimensional analysis (MDA), first introduced by Biber (1988), is widely-used framework for register analysis based on the theoretical assumption that register variation can be observed as differences in patterns of co-occurring lexico-grammatical features, which result from texts having register-specific communicative goals and contexts of use (Biber & Conrad 2019). It is a complex method that not only requires a reliable tagger for automatic identification of many linguistic features, but also entails numerous methodological decisions such as feature selection, identification of appropriate factors and establishing a factor loading threshold. However, whilst the recent emergence of freely available software tools (e.g., Nini 2019; Le Foll 2021) has made MDA more accessible to the research community, many MDA studies fail to report key parameters (Goulart & Wood 2021), which has a negative impact on transparency and replicability.

In this paper, we seek to examine the impact of factors that may influence the replicability of MDA studies. To this end, we present a comparative replication study that relies on three stratified random samples without replacement from the British National Corpus 2014 (Brezina, Hawtin & McEnery 2021; Love et al. 2017). Each 5-million-word sample comprises equal subsamples of the following registers: i) newspapers, ii) magazines, iii) fiction, iv) e-language and v) informal speech. The three BNC2014 samples were tagged with the MFTE (Le Foll 2021) and exploratory factor analyses were performed in R with various parameters. The resulting dimensions of variation and the dimensions scores of the various (sub-)registers were visualised and compared. This gives us the opportunity to discuss important methodological issues connected with the replication of MDAs including tagger performance across different registers, systematic vs. random variation and the impact of key methodological parameters.

**References**

Biber, Douglas. 1988. Variation across speech and writing. Cambridge: Cambridge University Press.

Biber, Douglas & Susan Conrad. 2019. Register, Genre, and Style. Second edition. Cambridge: Cambridge University Press.

Brezina, Vaclav, Abi Hawtin & Tony McEnery. 2021. The Written British National Corpus 2014 – design and comparability. Text & Talk 41(5–6). 595–615. <https://doi.org/10.1515/text-2020-0052>.

Goulart, Larissa & Margaret Wood. 2021. Methodological synthesis of research using multi-dimensional analysis. Journal of Research Design and Statistics in Linguistics and Communication Science 6(2). 107–137. <https://doi.org/10.1558/jrds.18454>.

Le Foll, Elen. 2021. Introducing the Multi-Feature Tagger of English (MFTE). Perl. <https://github.com/elenlefoll/MultiFeatureTaggerEnglish>.

Love, Robbie, Claire Dembry, Andrew Hardie, Vaclav Brezina & Tony McEnery. 2017. The Spoken BNC2014. International Journal of Corpus Linguistics 22(3). 319–344. <https://doi.org/10.1075/ijcl.22.3.02lov>.

Nini, Andrea. 2019. The Multi-Dimensional Analysis Tagger. In Tony Berber Sardinha & Marcia Veirano Pinto (eds.), Multi-Dimensional Analysis: Research Methods and Current Issues, 67–96. New York: Bloomsbury.

**Keywords:** multi-dimensional analysis (MDA), corpus methodology, replicability, replication, register variation, factor analysis, automatic annotation

[2453] ***Subverting racist discourse through parody: A CADS analysis of A.A Gill’s essay ‘Vegetarians’.***

Gill Philip (University of Macerata).

**Abstract.** In his essay ‘Vegetarians’, the journalist and food critic A.A. Gill takes aim at pandas, which he defines as ‘unique in the non-human mammalian world’ in having chosen to be vegetarian. The first part of this humorous essay draws heavily on the everyday language of ‘casual racism’, casting pandas as ungrateful immigrants supported by the State. This language is, however, subtly subverted through parody, hyperbole and irony, and in this study I aim to demonstrate both how the author effectively ‘defangs’ the racist discourse to cast derision on its users, and the potential dangers that are present when parody (and other humour) is misconstrued as serious discourse, especially when a sensitive topic is involved.

This study adopts a CADS approach, drawing on an ad-hoc corpus compiled from reader comments to Daily Mail articles on immigrants. The corpus features many well-worn tropes associated with migrants and which appear also in Gill’s text: that they are scroungers and parasites, accommodated and fed at the taxpayer’s expense; that they are work-shy and lazy; that they represent a danger to society. These tropes and their typical phraseological renderings as observed in the corpus are compared to the wordings used in the text to illustrate how the author simultaneously co-opts the discourse of petty racism and subverts it for humorous effect.

The findings are complemented by an analysis of multiple translations of the text and the comments made by students as they describe their attempts to echo equivalent racist discourses in their translations.

**Keywords:** CADS, anti-immigrant discourse, humour, translation

[2467] ***The prefixes anti- and pro- in Russian political discourse: A diachronic distributional semantics approach.***

Thomas Samuelsson (Stockholm University).

**Abstract.** This paper investigates the use of the antonymic prefix pair anti- ‘anti-’ and pro- ‘pro-’ in Russian political media in a time period that begins with the return of Putin to the Kremlin in 2012. Previous research has shown that the prefix anti- with the meaning of opposition and the prefix pro- with the meaning of support are both known to be active in Russian media (Zemskaya 1996). The prefix pro- has increased in activity (Ryazanova-Clarke and Wade 1999) and anti- is characteristic of contemporary Russian word-formation (Raciburskaja, Samyličeva and Šumilova 2015). In a British context, the prefix anti- has shown to be among the top keywords in British media and the collocates provide an overview over the analysed period in the news reporting (Tribble 2006).

Changes in the stems of the prefix anti- are observed to correlate with changes in the reporting about political and social topics in British newspapers (Duguid 2010).

In my paper presentation, I will present the results from the analysis of my own corpus with political language from the years 2012–2020. The data material consists of Russian political media texts from a diverse range of internet resources. The corpus, which contains 509 million words, is partitioned according to year and trained with the word embedding technique FastText. The produced semantic vector spaces are examined by studying the closest neighbours of the prefixes and the results are triangulated with extra-linguistic data. The investigation reveals diachronic changes in the closest neighbours of both anti- and pro-. The changes of anti- reflect dominating themes in the reporting during the actual time period, while the changes of pro- reveal themes of conflict and struggle in the Russian media reporting. The analysis illustrates how corpus linguistics can contribute to the research on morphology in political language.

**References**

Duguid, A. (2010). Investigating anti and some reflections on Modern Diachronic Corpus-Assisted Discourse Studies (MD-CADS). Corpora, 5(2), 191–220. <https://doi.org/10.3366/edinburgh/9780748640607.003.0005>

Raciburskaja, L. V., Samyličeva, N. A., & Šumilova A. V. (2015). Novye tendencii v sovremennom medijnom slovotvorčestve. In Novye tendencii v russkom jazyke načala XXI veka (pp. 134–221). Flinta.

Ryazanova-Clarke, L., & Wade, T. (1999). The Russian language today. Routledge.

Tribble, C. (2006). What counts in current journalism: keywords in newspaper reporting. In M. Scott, & C. Tribble (Eds.) Textual Patterns: Key Words and Corpus Analysis in Language Education (pp. 161–177). John Benjamins.

Zemskaja, E. A. (1996). Aktivnye processy sovremennogo slovoproizvodstva. In E. A. Zemskaja (Ed.) Russkij jazyk konca XX stoletija (1985–1995) (pp. 90–141). Jazyki russkoj kul’tury.

**Keywords:** corpus linguistics, distributional semantics, morphology, political language, Russian

[2520] ***Us investigate fragments today? Why why- and Mad-Magazine fragments in contemporary English?***

Yolanda Fernández-Pena (Universidade de Vigo) and Javier Pérez-Guerra (Universidade de Vigo).

**Abstract.** This paper focuses on two types of fragmentary constructions in English: why-fragments (1)-(2), and Mad Magazine sentences (3)-(4) (Akmajian 1984; Lambrecht 1990).

(1) Why deal with why-fragments?

(2) Why why-fragments?

(3) Me paint the house purple?

(4) [A: I heard you went out clubbing last Saturday.] B: Me go out clubbing?

Even though these fragmentary utterances can be equivalent in meaning to their fully-fledged sentential counterparts, the former may also convey an additional or ‘enriched’ nuance different from the ‘orthodox’ interpretation of the complete interrogatives. In fact, the reduced questions evoke a modal nuance equivalent to that of should (see Johnson 1975; Weir 2017; Zaitsu 2018, 2020) in (1) and (3), a uniqueness interpretation (Weir 2014) in (2) (‘why-fragments vs another type of fragment’) or serve to isolate the action from a specific time and question the possibility of such an action in general (Donaldson 2013: 13) in (4).

This paper explores the alternation between the canonical (i.e. equivalent to that of the corresponding complete sentences) and enriched interpretations of why- and Mad-Magazine constructions in contemporary English based on data from the BNC1994 DS (BNC Consortium 2007) and Spoken BNC2014 (Love et al. 2017) corpora. A binomial regression was used to determine which predictors (i.e. corpus/period, category of the fragments, and category mismatches between the fragment and their antecedents) are significant in explaining the semantics of the fragments.

To exemplify some of the results, our study shows that the timid (non-significant) increase in why-fragments revealed by the corpus data is supported by both higher probability of the so-called enriched interpretation of why-fragments in the Spoken BNC2014 data than in examples from BNC1994 DS, and the preference for novel expressions in the fragmentary construction (i.e. expressions not attested as such in the antecedent clauses) when the fragments have enriched meanings.

**References**

Akmajian, Adrian. 1984. Sentence types and the form-function fit. Natural Language & Linguistic Theory 2: 1–23.

Author. 2021

BNC Consortium. 2007. British National Corpus: XML edition. Oxford: Oxford Text Archive.

Donaldson, James. 2013. On elliptical why-questions. MSc dissertation. Edinburgh: The University of Edinburgh.

Johnson, David E. 1975. Why delete Tense? Linguistic Inquiry 6/3: 481–489.

Lambrecht, Knud. 1990. ‘What, me worry?’ – ‘Mad Magazine sentences’ revisited. In Proceedings of the Sixteenth Annual Meeting of the Berkeley Linguistics Society 16, 215–228.

Love, Robbie, Claire Dembry, Andrew Hardie, Vaclav Brezina and Tony McEnery. 2017. The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. International Journal of Corpus Linguistics 22/3: 319–344.

Weir, Andrew. 2014. Why-stripping targets Voice Phrase. Proceedings of NELS 43, 235–248.

Weir, Andrew. 2017. But write what? In Nicholas LaCara, Keir Moulton and Anne-Michelle Tessier eds. A Schrift to Fest Kyle Johnson. Amherst, MA: University of Massachusetts, 401–408.

Zaitsu, Anissa. 2018. Why make sense of silence? The clausal syntax of a reduced why-question. PhD, UC Santa Cruz.

Zaitsu, Anissa. 2020. Modality force and syntax in an understudied class of reduced why-questions in English. Glossa: A Journal of General Linguistics 5/1: 1–37.

**Keywords:** fragment, Mad Magazine sentences, British National Corpus, regression analysis, Construction Grammar

[2610] ***An intersectional corpus-assisted discourse study: The discursive construction of sexual misconduct in the Canadian Armed Forces.***

Rachelle Vessey (Carleton University).

**Abstract.** Sexual misconduct is a significant problem in the Canadian Armed Forces (CAF), with 13 senior Canadian military leaders (current and former) having been investigated since the beginning of 2021 and the term “sexual misconduct” itself identified as problematic and inadequate (Arbour, 2022, p. 14).

The Misconduct, Aggression, and Sexual Harassment (MASH) corpus consists of 1,676,293 words, drawn from 23,310 participants’ 47,330 short answer responses to eight separate surveys administered by the Canadian Department of National Defence (2014-2021). Using an intersectional (Crenshaw, 1991) corpus-assisted discourse studies (CADS) approach, this project examines how sexual misconduct is discursively constructed by CAF members according to their gender, rank, and (English/French) language and in explicit and implicit ways. The analysis of explicit discursive constructions focuses on differences in normalized frequencies, keyness, collocation, and concordance trends. The analysis of implicit constructions relies on an elaborated version of Coates and Wade’s (2007) framework for the Interactional and Discursive View of Violence and Resistance.

This paper presents an illustrative case study from the larger MASH project, using SketchEngine to examine the use of the words “harassment” and “misconduct” (and the French equivalents inconduite and harcèlement) according to gender, rank, and language. Findings suggest similarities and differences in terms of how (and the extent to which) sexual misconduct is labelled and discussed according to the individual positions accounted for within the corpus. With this research, the paper (1) responds to a call for CADS researchers to address intersectionality (Baker, 2020), (2) explicates how CADS can form part of a “clearly defined intersectional methodology” (Nash, 2008, p. 4), and (3) contributes to a growing body of intersectional CADS research (Bogetic, 2021; de la Ossa, 2019; Hunt and Jaworska, 2019; Flesch, 2020; Kitis, Milani and Levon, 2018).

**References**

Arbour, L. (2022). Report of the independent external comprehensive review of the Department of National Defence and the Canadian Armed Forces. Available: <https://www.canada.ca/en/department-national-defence/corporate/reports-publications/report-of-the-independent-external-comprehensive-review.html> (Accessed October 5, 2022).

Baker, P. (2020). Linguistic data and social groups. Data in discourse analysis (Technische Universität Darmstadt). Available: <https://www.linglit.tu-darmstadt.de/media/linglit/mitarbeitende/mueller_1/tagung_data_in_discourse_analysis/Baker.pdf> (Accessed October 5, 2022).

Bogetic, K. (2021). Co-opting the neoliberal manhood ideal: Masculinity, normativity, and recursive normalisation in Serbian gay men’s digital dating profiles. Language in Society, 50, 93-123.

Coates, L. and Wade, A. (2007) Language and violence: Analysis of four discursive operations. Journal of Family Violence, 22, 511-522.

Crenshaw, K. (1991). Mapping the margins: Intersectionality, identity politics, and violence against women of colour. Stanford Law Review, 43 (6), 1241-1299.

de la Ossa, A. C. (2019). Exceptionalising intersectionality: A corpus study of implied readership in guidance for survivors of domestic abuse. Gender and language, 13 (2), 224-250.

Flesch, M. (2020). lol thats how reddit talks ;) : Le site américain Reddit comme espace de variation de l’anglais. Étude de corpus intersectionnelle et quantitative d’usages non standard, au prisme du genre, de l’âge et de l’ethnicité. GLAD! (09, 2020). Available: https://journals.openedition.org/glad/2206 (Accessed October 5, 2022).

Hunt, S. and Jaworska, S. (2019). Intersections of nationality, gender, race and crime in news reporting: The case of Oscar Pistorius – Olympian and murderer. Discourse, Context & Media, 30, 100299, 1-11.

Kitis, E. D., Milani, T. M. & Levon, E. (2018). ‘Black diamonds’, ‘clever blacks’ and other metaphors: Constructing the black middle class in contemporary South African print media. Discourse & Communication, 12 (2), 149-170.

Nash, J. C. (2008). Re-thinking intersectionality. Feminist Review, 89, 1-15.

**Keywords:** corpus-assisted discourse studies, discourse, Canada, intersectionality, violence, military, sexual harassment

[2632] ***Discipline-specific lexical frames: Linguistic data extracted from corpora to inform EAP teaching across and within disciplines.***

Vivian Lameira (Universidade Federal do Rio Grande do Sul), Ana Eliza Bocorny (Universidade Federal do Rio Grande do Sul), Rozane Rebechi (Universidade Federal do Rio Grande do Sul), Deise Prina Dutra (Universidade Federal de Minas Gerais) and Tony Berber Sardinha (Pontifícia Universidade Católica de São Paulo).

**Abstract.** Lexical Frames (LFs), i.e. discontinuous sequences of words that form a structure (frame) around variable gaps (slots) (GRAY; BIBER, 2013), are building blocks of discourse, having, for this reason, great pedagogical value for the production of written academic genres. Little research has focused on the functional analysis of LFs in abstracts from a disciplinary and section-specific perspective. Here we investigate the frequency and functions of LFs used to linguistically realize rhetorical functions expressed in the sections of structured abstracts from four disciplines: (1) Computer and Information Sciences, (2) Physics, (3) Medicine and Health Sciences, and (4) Linguistics. In particular, the study seeks to identify and functionally analyze LFs using a data-driven approach. Data for the study were obtained from four corpora of abstracts, written in English and published in peer-reviewed journals. Each corpus, with one million words, was compiled with AntCorGen (ANTHONY, 2019) and analyzed with Sketch Engine (KILGARRIFF et al., 2004) and AntConc 4.0.10 (ANTHONY, 2022). A sample of the 200 most frequent LFs (50 from each discipline) reveals the existence of two major categories of discontinuous multiword units: (i) rhetorical LFs (RLF) and (ii) terminological LFs (TLF).

While the first are composed of lexical words indicating rhetorical functions across disciplines (presenting the aim of the study) and having in their composition elements that indicate rhetorical functions (aim, results, effects), the latter rarely occur across disciplines, as they refer to terms, procedures and concepts established in different disciplines (hazard ratio + + interval ci). As for the frequency of occurrence in the corpora, TLFs are less recurrent than RLFs, requiring lower cutoff points so that they can be extracted from the corpus. Finally, different from the RLFs, the TLFs have, in their composition, terms that indicate a link to a specialized area (for example, risk, hazard, patients).

**Keywords:** Corpus Linguistics, Discipline-specific lexical frames, Academic writing, Corpus-driven pedagogical resources

[2735] ***Upon my going into a coffee-house yesterday, and lending an ear to the next table…: A corpus-based exploration of Coffeehouse dialogues and their discursive practices in the period 1660-1712.***

Ersilia Incelli (Sapienza University).

**Abstract.** This research presents a corpus-based study which examines various speech-related written genres from the period 1660-1712, comprising transcribed coffeehouse dialogues, plays, poems, trial proceedings, as well as the written genre of broadsheets and newspapers (the Spectator

1712 and The Tatler 1709-1711), and extracts from diaries of prominent scholars such as Robert Hooke and Samuel Pepys. The corpora and subcorpora of collected texts reflect the popularity of public coffeehouses in England, renowned as social spaces where people gathered news and debated ideas on politics, religion, science, poetry, literature, travel and other matters. Although coffeehouses have been studied from the historical, social theorist point of view (Habermas 1989), this research adds linguistic insight into the experience of these public spaces, revealing the peculiarity of changing relationships between self and society. Hence, the aim of the study is to explore the linguistic features relevant to the sociohistorical and pragmatic aspects of the texts in their sociocultural, communicative contexts, focusing on speech acts, politeness strategies, conversation principles, hedging and assertive utterances. Data retrieved so far reveal a significant tension in the supposed discursive practices. The idea of the coffeehouse as a peaceful place of rational, genteel discourse, egalitarian dialogue and reasoned argumentation (Habermas 1989), contrasts sharply with linguistic evidence which reveals coffeehouse conversations as not always civil (you Rogue), and orderly (On this a quarrel soon began), with unrestrained turn taking and frequent impoliteness (But parted each to shew his spight, Curses, instead of saying- Good Night; Sir,… I’d kick him headlong to the Devil). This study also looks at some of the recurrent pragma-dialectical features such as: high frequency pronouns I, thou; referencing style, e.g. sir, discourse markers e.g. pray, nay.

**Keywords:** coffeehouses, discursive practices, historical (im)politeness, public sphere, pragma-dialectical features, specialized discourse community

[2781] ***National collections as corpora: Insights from the Our Heritage, Our Stories project.***

Ewan Hannaford (University of Glasgow) and Marc Alexander (University of Glasgow).

**Abstract.** Institutional and, more broadly, national collections can be invaluable resources for corpus linguistic inquiry. Ideally, these bodies offer ready access to large quantities of thematically coherent data from a diverse range of sources, enabling analysis that strongly adheres to the key corpus principles of representativeness, sampling and balance (Baker & McEnery, 2015). However, these collections are often focused on, and curated, in mainstream language varieties (Seifart et al., 2018), ignoring materials reflecting regional and social language varieties. This is particularly true in the case of minority groups and other under-represented communities in the UK, whose cultural contributions have typically been side-lined by archival institutions in favour of ‘treasured’ material documenting mainstream populations, narratives, and linguistic varieties (Prescott & Hughes, 2018). Consequently, the linguistic representativeness of institutional collections in the UK, and, by extension, the UK national collection, is currently limited, hindering corpus linguistic research on regional and social language varieties and inhibiting broader comparisons between these varieties and mainstream materials.

The AHRC-funded Our Heritage, Our Stories project aims to reduce this shortcoming by integrating diverse community-generated digital content into the UK national collection, and the proposed talk discusses work on this project as a central case study. After first addressing the rationale and approach of this work, we will discuss the conceptual, ethical, and technical challenges involved in situating and promoting regional and social language varieties alongside existing standardised materials, without flattening the linguistic complexity and diversity of said materials. Whilst discussing approaches to democratising production/curation of materials containing regional and social language varieties and facilitating their integration into more representative collections, we will also exemplify new avenues of corpus linguistic research that will be opened by these resources, highlighting the potential of linguistically diverse national collections as transformative tools for language equity, language preservation, and corpus studies.

**References**

Baker, P. & McEnery, T. (2015). Introduction. In P. Baker & T. McEnery (eds.), Corpora and discourse studies (pp. 1–19). Palgrave Macmillan. <https://doi.org/10.1057/9781137431738>

Prescott, A. & Hughes, L. M. (2018). Why do we digitize? The case for slow digitization. Archive Journal, available at: <https://www.archivejournal.net/essays/why-do-we-digitize-the-case-for-slow-digitization/>

Seifart, F., Evans, N., Hammarström, H., & Levinson, S. (2018). Language documentation twenty-five years on. Language, 94(4), 324-345. <https://doi.org/10.1353/lan.2018.0070>

**Keywords:** corpora, archives, language varieties, language preservation

[2806] ***The effect of speaking tasks on lexical complexity and fluency: Using corpus methods to inform language assessment.***

Emil Tangham (Lancaster University).

**Abstract.** Adapting language to interactional contexts is an essential component of effective L2 communication (Biber & Conrad, 2019). L2 proficiency assessments such as IELTS and APTIS use different tasks to elicit test-takers’ production in order to evaluate their proficiency (British Council, 2020), but typically do not examine how learner language changes across tasks. This study investigates changes in learner production across four tasks in a spoken assessment and the extent to which this process is connected to L2 proficiency. The findings have applications in language testing, teaching, and SLA research. Specifically, the study addresses the following research questions:

• What is the effect of speaking task on lexical complexity and fluency in test-takers’ production?

• Is complexity and fluency of L2 speech in different tasks affected by L2 proficiency?

The study employs the spoken part of the British Council-Lancaster Aptis Corpus, which is based on the British Council Aptis General Test, a global English proficiency examination (British Council, 2020). The corpus contains 663,111 words from 865 speakers at three levels of proficiency (CEFR: B1/B2/C). To measure lexical complexity, lexical diversity (MATTR, Guiraud’s Index) and density measures (content/total word ratio) were used; filled and unfilled pause frequency was calculated to establish speakers’ fluency. The corpus was analysed using #LancsBox (Brezina et al., 2021) and Sketch Engine (Kilgarriff et al., 2014), using repeated-measures ANOVA to compare test-taker’s production across tasks and at different levels of proficiency.

The results showed that the task effect was observed both, with respect to fluency and lexical complexity. The most notable interaction between L2 proficiency and task performance was found with respect to unfilled pause frequency and lexical density. The findings will be discussed in the context of previous task-based L2 studies (e.g. Gablasova et al., 2017; Yan et al., 2021), language testing, and test validation.

**References**

Biber, D., & Conrad, S. (2019). Register, genre, and style (Second edition). Cambridge, UK: Cambridge University Press. <https://doi.org/10.1017/9781108686136>

Brezina, V., Weill-Tessier, P., & McEnery, A. (2021). #LancsBox v. 6.x. [software package]

British Council (2020) Aptis General Candidate Guide. [https://www.britishcouncil.org/sites/default/files/aptis\_candidate\_guide\_2020\_a4 \_final.pdf](https://www.britishcouncil.org/sites/default/files/aptis_candidate_guide_2020_a4%20_final.pdf)

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. <https://doi.org/10.1093/applin/amv055>

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: ten years on. Lexicography, 1(1), 7–36. <https://doi.org/10.1007/s40607-014-0009-9>

Yan, X., Kim, H. R., & Kim, J. Y. (2021). Dimensionality of speech fluency: Examining the relationships among complexity, accuracy, and fluency (CAF) features of speaking performances on the Aptis test. Language Testing, 38(4), 485–510. <https://doi.org/10.1177/0265532220951508>

**Keywords:** Communicative Competence, Language Assessment, Test Validation, Task-Based, SLA

[2833] ***A new model of variation in research articles: Optimizing multi-dimensional analysis to focus on disciplinary variation.***

Bethany Gray (Iowa State University) and Paul Thompson (University of Birmingham).

**Abstract.** Multi-dimensional (MD) analysis (Biber 1988) results in functionally-motivated linguistic descriptions of the domains being analyzed, based on the identification of statistical co-occurrence patterns of large sets of linguistic features. Although developed to study variation between broad register categories, MD analysis has also been applied productively to more restricted domains, such as variation within academic writing (e.g., Gardner et al. 2019, Goulart 2021 on student writing; Egbert 2015 on textbooks, popular academic texts, and journal articles; Gray 2015 on disciplinary variation in research articles; Thompson & Hunston 2020 on interdisciplinary discourse). However, many previous MD analyses of academic sub-registers have relied on the same features as used in MD analyses of general registers, unlike research on other specialized registers which have extended to include register-specific features (Friginal 2008). In addition, corpora used for MD analyses of disciplinary variation have been restricted in size and disciplines represented.

The present study addresses these gaps through a new multi-dimensional analysis of disciplinary variation using the ELSE49 corpus, a sample of 14,204 research articles (c. 72 million words) in 10 social science and science disciplines (Management, Economics, Linguistics, Environmental Science, Agricultural Science, Animal Science, Biology, Materials Science, Medicine, Physics).

A new MD model is carried out using an expanded feature set, including traditional MD features from Biber’s Tagcount program (e.g., short passives, past tense), features characteristic of academic writing (e.g., noun+prep, existential there, exemplification markers, demonstratives + shell noun), and a new set of 51 semantic categories (e.g., research nouns, calculation verbs) developed specifically for academic writing research (Authors 2022). In this presentation, we present the resulting 5-dimensional model (variance explained = 32.4%), with a particular focus on Dimension 1 (‘Elaboration and Evaluation vs. Procedural Narration’, R2 = .514) and Dimension 2 (‘Human and Communication Focus vs. Quantitative Discourse’, R2 = .570).

**Keywords:** multi-dimensional analysis, disciplinary variation, academic writing, textual variation

[2874] ***Multidimensional analysis of syntactic complexity development in L2 learner writing in a university IEP programme in the U.S.***

Sangeun Kim (University of Exeter) and Joyce Lim (Aston University).

**Abstract.** Recent research findings inform phrasal syntactic complexity features as a significant predictor of English writing quality (e.g., Larsson & Kataari 2020; Kyle & Crossley 2018). However, longitudinal studies using L2 written texts produced in a natural classroom setting have been relatively rare.

The current study aims to add longitudinal evidence of English syntactic development with a focus on co-occurring patterns in syntactic measures in academic writing produced by speakers of English as an additional language (L2) over time.

The longitudinal data comes from the University of Pittsburgh English Language Institute corpus (PELIC; Juffs et al. 2020), where the written samples were collected from the same participants over three consecutive academic semesters in writing classes. This study adopts a multidimensional (MD) analysis (Biber, 1988). A comprehensive framework of 71 linguistic features is established to reflect the two syntactic dimensions: the structural form and syntactic functions (Biber et al., 2021, pp.14-15), consisting of phrasal and clausal features. A factor analysis was conducted to identify emerging functional dimensions. Subsequently, a mixed-effect model is fitted to rule out the random effects of individual students and measure mediating effects of English level and first language background interacting with the 'time' variable.

The findings confirm the previously noted spoken and written register distinction (evidence of spoken register in novice academic writers) as an indicator of academic English development in L2 writing (e.g., Biber & Gray 2013; Kobayashi & Abe 2016; Kim & Nam 2019). As the students have become more proficient, the learner texts demonstrated a closer association with phrasal complexity associated with more abstract and informational topics. The overall syntactic variation indicates the significant influence of communicative purposes actively chosen by the writer, which invites further investigation into the interaction between linguistic, writer and situational variables.

**References**

Biber, D. (1988). Variation across Speech and Writing. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511621024

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development?. Tesol Quarterly, 45(1), 5-35.

Biber, D., & Gray, B. (2013). Discourse characteristics of writing and speaking task types on the TOEFL ibt® test: a lexico‐grammatical analysis. ETS Research Report Series, 2013(1), i-128.

Biber, D., Reppen, R., Staples, S., & Egbert, J. (2020). Exploring the longitudinal development of grammatical complexity in the disciplinary writing of L2-English university students. International Journal of Learner Corpus Research, 6(1), 38-71.

Gray, B., Geluso, J., & Nguyen, P. (2019). The longitudinal development of grammatical complexity at the phrasal and clausal levels in spoken and written responses to the TOEFL iBT® test. ETS Research Report Series, 2019(1), 1-51.

Juffs, A., Han, N-R., & Naismith, B. (2020). The University of Pittsburgh English Language Corpus (PELIC) [Data set]. <http://doi.org/10.5281/zenodo.3991977>. Retrieved from: <https://eli-data-mining-group.github.io/Pitt-ELI-Corpus>. Accessed 24 June 2021.

Kim, J. E., & Nam, H. (2019). How do textual features of L2 argumentative essays differ across proficiency levels? A multidimensional cross-sectional study. Reading and Writing, 32(9), 2251-2279.

Kobayashi, Y., & Abe, M. (2016). A Corpus-Based Approach to the Register Awareness of Asian Learners of English. Journal of Pan-Pacific Association of Applied Linguistics, 20(2), 1-17.

Kyle, K., & Crossley, S. A. (2018). Measuring Syntactic Complexity in L2 Writing Using Fine-Grained Clausal and Phrasal Indices. The Modern Language Journal, 102(2), 333-349. <https://doi.org/10.1111/modl.12468>

Larsson, T., & Kaatari, H. (2020). Syntactic complexity across registers: Investigating (in)formality in second-language writing. Journal of English for Academic Purposes, 45, 100850. <https://doi.org/10.1016/j.jeap.2020.100850>

**Keywords:** multidimensional analysis, L2 longitudinal study, syntactic complexity development, L2 student register variation, phrasal and clausal complexity

[2928] ***The verbal language of visual arts: a multi-dimensional analysis of the discourse on Sally Mann's photography.***

Yara Maria de Toledo Dias Romeiro (PUC-SP).

**Abstract.** The visual arts have a profound effect on society, yet there has been little interest in the language used to talk about the visual arts in Linguistics. This is in contrast to its economic impact. In 2013, the arts industry generated US$ 704.6 billion, 4.2% of the US GDP (cf. Gifford, 2016). Texts about the visual arts perform important functions in the field, by supporting, highlighting, describing, assessing, educating about, and theorizing about the artistic object itself (cf. Lazzeretti, 2016). Based on such discourse, art theories are formed and concepts through which art is regarded are developed. The objective of this research is to investigate this discursive universe of photography from a Lexical Multidimensional Analysis perspective (Berber Sardinha, 2014, 2017, 2019), which is a statistically-based approach capable of detecting groups of lexical items that co-occur in texts revealing recurrent themes in the corpus. To conduct this research, a corpus consisting of 555 texts (nearly 765.000 words) representing the discourse about the US photographer Sally Mann was collected and analyzed. The corpus comprises 12 different registers, ranging from her books to the wall texts, to articles in the press, among others. The resulting seven factors were interpreted as lexical dimensions related to the themes underlying Mann’s photographic work, centered on her hometown, the fascination with mortality, and family relations. These texts were sourced from Mann’s official library and public depositories. This diachronic corpus covers 30 years. The texts were cleaned up, then lemmatized and tagged for part-of-speech. A script captured the major lemmas, counted, and normed their occurrences. The normed counts were submitted to a factorial analysis. The factors were interpreted for the underlying discourses. This research intends to contribute to the use of empirical, corpus-based linguistic studies as a tool to investigate the realm of visual arts and their discourses.

**Keywords:** Lexical Multidimensional Analysis, Corpus Linguistics, Discourse Analysis, Visual Arts, Photography

[2943] ***Exploring the predictive power of multiple lexical diversity measures for L2 Spanish writing proficiency.***

Earl Brown (Brigham Young University), Brett Hashimoto (Brigham Young University) and Alan Brown (University of Kentucky).

**Abstract.** Researchers (Gebril & Plakans, 2016; Crossley & McNamara, 2012) have noted a relationship between lexical diversity (LD) and writing proficiency, accounting for as much as 44% of variability in writing proficiency (Crossley et al., 2011). However, it remains uncertain which LD measure, or measures, best predict writing proficiency (e.g., McCarthy & Jarvis, 2010). Recently, Woods et al. (2023) utilized Principal Components Analysis (PCA) and mixed-effects linear regression to identify which combination of LD measures best accounted for writing proficiency in a corpus of L2 English writing. They found that a statistical model with multiple LD measures better predicted writing proficiency than models with only one LD measure. The current study adopts a methodology similar to that of Woods et al., but with L2 Spanish. The Corpus Escrito del Español L2 'Written Corpus of L2 Spanish' was accessed and two Python modules (i.e., lexical-diversity [Kyle, 2020] and lexicalrichness [Shen, 2022]) were utilized to calculate 12 LD measures. A total of 1,195 writing samples of at least 50 words composed by 843 women and 352 men, ranging in age from 13 to 88 years old (mean = 25.0, median = 20, SD = 11.9) were subjected to analysis. A series of mixed-effects linear regression models showed that several LD measures best predict proficiency in these data, and that the first two principal components capture different information about the LDs that they are based on. The results highlight the importance of using several LD measures rather than relying on one when assessing L2 proficiency.

**Keywords:** Lexical diversity, L2 Spanish, Principal Components Analysis, Mixed-effects linear regression

[2987] ***Arabic Metaphor Corpus (AMC) with Semantic and Sentiment Annotation.***

Israa Alsiyat (Lancaster University), Scott Piao (Lancaster University) and Mansour Almansour (Lancaster University).

**Abstract.** Metaphor is an inseparable part of communication in Arabic language. Over the past years, Arabic metaphors have been increasingly used in online communication, which has drawn attention of researchers. Arabic metaphors in online context have been evolving and show distinct features, reflecting the need for speedy online communication compared to traditional written forms. While Arabic metaphors have intrinsic ambiguity of word sense, the metaphors used online present an even higher challenges for accurate interpretation of their true meanings in their contexts. For example, no reliable resources can be found for online variants of Arabic metaphors for interpretation, newly emerging metaphor structures are not registered in standard Arabic literature, they can have different writing styles in different Arabic dialects, and frequently appearing typos etc. Therefore, we need a corpus that reflects online Arabic metaphor usage annotated with their key features for the investigation of Arabic metaphor usage in online context and to develop tools for identifying and classifying them. However, as far as we are aware, there is no such corpus publicly available yet.

To address this issue, we have built an Arabic Metaphor Corpus (AMC), which contains 1,000 online book reviews from LARB Corpus (Aly and Atiya, 2013). The AMC was manually annotated by two native Arabic annotators for sentiment of reviews and metaphors, then we merge different sets of annotation into gold standard annotations. Furthermore, the reviews were tagged using the UCREL Arabic Semantic Tagger AraSAS (El-Haj et al., 2022) to facilitate detailed analysis of semantic patterns and contexts of the metaphors.

The AMC is available online (<https://github.com/IsraaMousa/ArabicMetaphorCorpus/blob/main/Arabic-Metaphor-Corpus.xlsx>). The AMC corpus is a unique resource for the study of Arabic metaphors in online review context, and will contribute a valuable corpus resource to the corpus linguistics community. The average length of the reviews is 107.968 tokens, and the longest and shortest reviews have 1,084 and 2 tokens respectively. In the AMC, according to our gold annotation, there are 702 and 171 positive and negative reviews, making up 70.2% and 17.1% of the reviews respectively. The remaining 127 reviews are judged to be neutral. In regard to the metaphors annotated in the AMC, 602 and 272 metaphors are manually classified as positive and negative ones respectively. The remaining 126 metaphors are classified as neutral.

In this presentation (see the presentation slides), we report the data selection criteria, methodology of manual annotation, and statistical information of data structure and annotation. We will also discuss challenging issues arisen during the corpus construction process.

**References**

Aly, M. and A. Atiya (2013). Labr: A large scale arabic book reviews dataset. In Proceedings of The 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria. Volume 2: Short Papers. pp. 494–498.

LABR corpus is available at website: <https://scholar.googleusercontent.com/scholar?q=cache:0S-Kb9VPnocJ:scholar.google.com/+large+Arabic+book+reviews&hl=en&as_sdt=0,5> (accessed May 29, 2023).

El-Haj, M., E. de Souza, N. Khallaf, P. Rayson, and N. Habash (2022). AraSAS: The Open Source Arabic Semantic Tagger. In Proceedings of The 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur’an QA and Fine-Grained Hate Speech Detection. Marseille, France, pp. 23–31.

**Keywords:** Arabic Metaphor Corpus, Sentiment Analysis, Metaphor, Arabic Corpus Linguistics, Arabic NLP

[3002] ***“O women, give charity, for verily I was shown that you make up the majority of the people of Hellfire”: Using Transitivity to explore the Islamic State through the lens of the female propagandist writer.***

Katie Patterson (University of Granada), Miguel Ángel Benitez Castro (University of Zaragoza) and Encarnación Hidalgo Tenorio (University of Granada).

**Abstract.** Whilst not sharing the freedoms pertaining to men, women perform a variety of active roles within the Islamic State. Apart from producing offspring who will support and defend it, they also serve as ideological educators, operational facilitators, deadly fighters and emotional and financial support providers for male relatives (Ali, 2006). Previous linguistic studies have focused on the traditional role of the female warriors or mujahidaat (Daniel, 2022), or on the terms that define both the subjectivity and the lack of agency of these women against their husbands or children (e.g., jihadi bride, ISIS widows, or ISIS mother) (cf. Martini, 2018; French, 2020). However, little has been documented of the propaganda steered specifically towards women by women. Our study explores a number of questions regarding other-presentation, self-perception and identity from the perspective of the female jihadi writer, as the conduit for inspiring, motivating and recruiting other females. Through a Transitivity analysis (see Halliday and Matthiessen, 2014) of social actors evident in our corpus of propaganda magazine entries (Dhiab-Hassan, Benítez-Castro and Hidalgo-Tenorio, 2018), the research hopes to answer four questions:

1. How do female propaganda writers of IS construe themselves and others?

2. To what extent do the writers show self-agency?

3. Does the type of article (indoctrination versus recruitment) influence how they construe themselves and others?

4. Do notions of identity or self-agency develop through the evolution of Islamic State (2015-2017)?

For our analysis, we will combine both quantitative and qualitative approaches, and will employ O’Donnell’s (2018) UAM CorpusTool tagging software programme. The paper aims to help deconstruct the simplistic status of gender and discourse in relation to us vs. them that is present in Western societies (Martini, 2018). We argue for a multi-causal and multi-faceted approach to female radicalisation, ultimately hoping to extend rather than to limit understanding of these women’s actions and self-representation.

**References**

Ali, F. 2006. Rocking the cradle to rocking the world: The role of Muslim female fighters. Journal of International Women’s Studies, 8(1): 21-35.

Daniel, L. 2022. tactical mistakes: Women’s role in terrorism. Liberty University Journal of Statesmanship & Public Policy, 2(2): 1-9.

Dhiab-Hassan, A., Benítez-Castro, M. Á. and E. Hidalgo-Tenorio. 2018. Nutcracker: The JIHAD Corpus. University of Granada.

French, N. S. 2020. Mujahidin, mujahidat: balancing gender in the struggle of Jihadi-Salafis. In The Routledge Handbook of Islam and Gender, ed. by J. Howe, (pp. 269-281). London: Routledge.

Halliday, M. A. K. and Christian M. I. M.. 2014. An Introduction to Functional Grammar. New York: Oxford University Press.

Martini, A. 2018. Making women terrorists into “jihadi brides”: An analysis of media narratives on women joining ISIS. Critical Studies on Terrorism, 11(3): 458-477.

O’Donnell, M. 2018. UAM CorpusTool. <http://www.corpustool.com/>.

**Keywords:** Corpus Linguistics, Critical Discourse Analysis, Transitvity, Islamic State, Gender

[3038] ***Can low-proficiency level learners produce diverse texts? A multidimensional approach to Czech as a foreign language.***

Adrian Jan Zasina (Institute of Czech Studies, Charles University).

**Abstract.** Many recent studies (Biber et al., 2016; Friginal & Weigle, 2014; Staples et al., 2018; Weigle & Friginal, 2015) predominantly focus on the writing of advanced learners. Written production of beginners is mostly analysed in comparison with those from other proficiency levels (Kim & Nam, 2019). However, particular attention has not been paid to text variability in low-level written production. This study thus attempts to analyse functional variability within texts of Polish learners of Czech at level A2 (basic users).

Using the Czech model of multidimensional analysis (Cvrček et al., 2018), learner texts (62 texts from 16 students) were projected onto a multidimensional space to identify their variability. The texts were divided into 4 tasks including: an informal letter, a description of a place, an argumentative essay, and a story. The results indicate that the writing of beginner learners reveals differences between the tasks, which underlines the usefulness of the tasks in the comparison of L2 writing; letters are characterised by lowest cohesion and the highest amount of addressing coding, descriptions of place are spontaneous and polythematic, argumentative essays are the most cohesive, with the lowest amount of addressee coding among all tasks, and stories describe the most concrete, particular information and are highly progressive. Moreover, within-task variability showed that argumentative essays have the lowest degree of variation, as they are more convention-based than stories, which have the highest degree of variation. Comparison with the target Czech registers showed that non-native argumentative essays are close to native argumentation; however, they differ in their use of verbal phrases that are more frequent in learner writing. Storytelling, however, represents a task that has no close counterpart in the native Czech data. The vaguest, in terms of Czech register, are informal letters and place descriptions.

**References**

Biber, D., Gray, B., & Staples, S. (2016). Predicting Patterns of Grammatical Complexity Across Language Exam Task Types and Proficiency Levels. Applied Linguistics, 37(5), 639–668. <https://doi.org/10.1093/applin/amu059>

Cvrček, V., Komrsková, Z., Lukeš, D., Poukarová, P., Řehořková, A., & Zasina, A. J. (2018). Variabilita češtiny: Multidimenzionální analýza. Slovo a slovesnost, 79(4), 293–321.

Friginal, E., & Weigle, S. (2014). Exploring multiple profiles of L2 writing using multi-dimensional analysis. Journal of Second Language Writing, 26, 80–95.

Kim, J., & Nam, H. (2019). How do textual features of L2 argumentative essays differ across proficiency levels? A multidimensional cross-sectional study. Reading and Writing, 32(9), 2251–2279. <https://doi.org/10.1007/s11145-019-09947-6>

Staples, S., Biber, D., & Reppen, R. (2018). Using Corpus-Based Register Analysis to Explore the Authenticity of High-Stakes Language Exams: A Register Comparison of TOEFL iBT and Disciplinary Writing Tasks. The Modern Language Journal, 102(2), 310–332. <https://doi.org/10.1111/modl.12465>

Weigle, S. C., & Friginal, E. (2015). Linguistic dimensions of impromptu test essays compared with successful student disciplinary writing: Effects of language background, topic, and L2 proficiency. Journal of English for Academic Purposes, 18, 25–39. <https://doi.org/10.1016/j.jeap.2015.03.006>

**Keywords:** Czech, learner corpus, multidimensional analysis, textual variation, register variability, SLA, writing skill

[3056] ***‘Windows and mirrors’ on ill health in best-selling young adult fiction.***

Maria Leedham (The Open University), Sally Hunt (The Open University) and Sarah Mukherjee (The Open University).

**Abstract.** Children’s literature offers ‘windows and mirrors’ onto the world (Bishop 1990); that is, books for children and young adults provide other world views and experiences (windows) and also self-reflections (mirrors). This paper describes a corpus-assisted discourse studies investigation of the ‘windows and mirrors’ on ill health within a corpus of the 50 most commercially-successful books for young adults (11-18 years) sold in the UK over a 5-year period (2017-2022).

The mental health of many young people has been substantially impacted by the COVID-19 pandemic (UK Gov), in addition to the difficulties presented by physical illness. Books act as resources for young adults, allowing readers to build conceptions of illness and ways of coping. This paper reports on patterns found in the representations of illness and in characters’ responses. The language of illness and resilience in young people’s reading repertoires is likely to have an impact on their understandings of their own lives and the ways in which they construct their own responses to ill health.

Our analysis involves the comparison of a sub-corpus of books dealing with health issues (both mental and physical) to the whole corpus of c.5 million words. It reveals interesting correlations between character gender and process type, with mental processes, i.e. those relating to thinking and feeling, being especially pertinent. Our understanding of the patterns and their meanings is enriched by the insights of secondary school students with whom we have shared our findings in focus groups.

Key findings from the study will be disseminated to stakeholders including young people, English teachers and school librarians. We aim to provide them with the ‘windows and mirrors’ - or constructions - of ill health that young people encounter in their reading and so to recognise and understand relevant coping strategies they may adopt.

**Keywords:** mental health, physical health, young adult fiction, education, secondary school, literature, CADS

[3058] ***Between Individuation and Argumentation: “A Few” vs. “Several”.***

Nicole Katzir (Tel Aviv University).

**Abstract.** The present study investigates the differences between the quantifiers "a few" and "several". Previous research has shown that "several" refers to slightly higher quantities than "a few" (e.g., Newstead & Coventry, 2000). However, this study argues that the difference between the two quantifiers goes beyond quantity. Namely, that these quantifiers construe the entities they introduce differently, such that "several" construes its nominal complement as being composed of individuated entities. Individuated entities are perceived as more prominent. Interlocutors pay more attention to prominent entities; therefore, they are perceived as important and carry heavier argumentative weight (as defined by Anscombre & Ducrot, 1976).

Using data from COCA (Davies, 2008-), two Distinctive Collexeme Analyses (Gries & Stefanowitsch, 2004) were conducted to identify the head nouns and adjectives that distinguish between the quantifiers in the determiner construction (several/a few [adj different] [n factors]).

The results show that nouns which are rejected by the partitive construction are more likely to be distinctive for "a few". Composed of a numeral/quantifier, "of", and a noun, this construction carves out a new discourse referent out of an existing one. When a plural entity is composed of individuated sub-parts, these sub-parts are more prominent, and thus make for better candidates to serve as discourse referents.

For adjectives, adjectives that convey prominence are more likely to be distinctive for "several", while adjectives that convey insignificance are more likely to be distinctive for "a few". These findings suggest that nominals modified by several tend to have greater discourse prominence, making the entire nominal argumentatively stronger.

Additionally, a modified DCA of constructions and adverbs that mark different aspects of argumentativity revealed that "a few" is associated with weak argumentativity, while "several" is associated with strong argumentativity. Taken together, these findings indicate that the difference between the quantifiers isn't solely due to quantity.

**References**

Anscombre, J.-C., & Ducrot, O. (1976). L’argumentation dans la langue. Langages, 42, 5–27.

Davies, M. (2008-). The Corpus of Contemporary American English (COCA): 600 million words, 1990-present. <https://www.english-corpora.org/coca/>

Gries, S. T., & Stefanowitsch, A. (2004). Extending collostructional analysis. International Journal of Corpus Linguistics, 9(1), 97–129.

Newstead, S. E., & Coventry, K. R. (2000). The role of context and functionality in the interpretation of quantifiers. European Journal of Cognitive Psychology, 12(2), 243–259. <https://doi.org/10.1080/095414400382145>

**Keywords:** Quantifiers, Argumentation in language, Discourse prominence, Distinctive Collexeme Analysis

[3075] ***Authority and visibility in L2 (Czech) learner academic discourse: Realisations, functions and distribution of self-mention across master’s theses.***

Olga Dontcheva-Navratilova (Masaryk University).

**Abstract.** This contribution explores self-mention in L2 (Czech) English-medium master’s theses written in the humanities along three dimensions of analysis: realisation, authorial roles and distribution across rhetorical sections. While extending the scope of self-mention to cover nominal forms, the aim of the study is to find out how Czech graduates combine pronominal and nominal self-mention to modulate the degree of visibility and authority they convey in their texts. The contrastive corpus-based investigation compares a corpus of Czech English-medium master’s theses in linguistic, literary studies and ELT methodology with two reference L1 corpora representing learner (a selection from BAWE) and published academic discourse (research articles) in the same fields to examine differences pertaining to variation along culture, expertise and discipline dimensions. SketchEngine was used for searching for occurrences of pronominal and nominal self-mention.

The findings indicate that realisation patterns of self-mention and preferences for specific authorial roles vary significantly across the corpora. In comparison to L1 expert and learner writers, Czech students underuse pronominal self-mention and overuse nominal self-mention. This deviation from L1 academic writing conventions may reflect the students’ efforts to find a balance between the need to claim credibility and express humility. On the level of rhetorical structure, Czech graduates show an awareness of genre conventions, yet their choices diverge from disciplinary patterns in published research articles. A preference for less powerful authorial roles suggests that Czech students are reluctant to assume an authoritative stance, thus confirming that they are still developing their rhetorical skills. This may be motivated by the efforts of students to blend L1 and L2 academic conventions, their lower level of rhetorical maturity, and the audience addressed in an examination context.

**Keywords:** self-mention, master's thesis, author roles, metadiscourse, academic discourse, intercultural rhetoric

[3163] ***How expert are universities? Investigating research mobilities in primary literacy education in England.***

Anna Lindroos Čermáková (Lancaster University) and Julia Gillen (Lancaster University).

**Abstract.** ‘Research mobilities in primary literacy education in England’ project investigates the movements of literacy research to and among primary school teachers in England in a shifting and politically charged context. We explore why some kinds of literacy research are gaining traction while others are not and who are the social actors of significance. We focus on how expertise is

constructed as researchers find themselves on a competitive market with other powerful stakeholders (Puschmann 2015). We use two purpose-built corpora: UK newspapers corpus (ca. 415,000 words; 2017-2021) and a Twitter corpus (ca. 210,000 words/5,500 tweets; 2019-2021). With keyword analysis we explore how the expertise discourses differ in these two corpora. The expertise in the newspaper corpus is constructed around lexis we were anticipating, e.g., research, science, evidence, university; however, this vocabulary is, surprisingly, virtually missing in our Twitter data. Based on this analysis, we further interrogate our corpora in a complementary, rather than comparative way, and we ask the following research questions: 1) How is expert knowledge originating from universities constructed in the newspaper corpus? and 2) Which academic Twitter accounts are “influential” and how is this influence exercised?

The verb collocate analysis of ‘university’ reveals various types of research attribution (or frequently non-attribution) (Hunston 2010, 2015), including commitment to the statement (suggests), results/process (e.g., found vs conducted), or evaluation (warns). Mapping the network between university mentions and affiliated researchers and media outlets shows which universities figure prominently in which newspapers, e.g., University of Dundee being the most frequently mentioned university by several newspapers or Lancaster University being mentioned only in the Times Educational Supplement. In the Twitter data, we operationalise the “influence” through number of retweets and account followers and identify the most “influential” accounts with some surprising results, e.g., @matthewcobb, the account of a Professor of Zoology.

**References**

Hunston, S. (2010). Corpus Approaches to Evaluation: Phraseology and Evaluative Language. London: Routledge.

Hunston, S. (2015). Talking Science. Science in the news on BBC Radio. In M. Bondi et al. (Eds), Discourse in and Through Media. Recontextualising and Reconceptualising Expert Discourse, pp. 66-91. Cambridge: Cambridge Scholars Publishing.

Puschmann, C. (2015). A Digital Mob in the Ivory Tower? Context collapse in the scholarly communication online. In M. Bondi et al. (Eds), Discourse in and Through Media. Recontextualising and Reconceptualising Expert Discourse, pp. 22-45. Cambridge: Cambridge Scholars Publishing.

**Keywords:** literacy, construction of expertise, research mobility, media discourse, Twitter

[3294] ***The semantic prosodies of different senses of a word.***

Mathias Russnes (University of Oslo).

**Abstract.** In previous research on semantic prosody, lexical items with multiple senses have predominantly been treated as monosemous (e.g. Partington, 1998; Stubbs, 2002). Building on Bublitz (1996), this paper challenges this and aims to display that polysemy should be considered when establishing the prosody of lexical units, specifically addressing the research question: To what extent is the core word’s meaning(s) decisive for the semantic prosody of lexical units? To answer this, a corpus study of two verb lemmas previously ascribed negative prosodies is performed, namely COMMIT and HAPPEN (see Stewart, 2010), where each sense is analysed after Sinclair’s (1996) extended-unit-of-meaning-model. Drawing on material from the British National Corpus, the study demonstrates that the separate senses of the core of an extended unit of meaning can trigger markedly different prosodies.

For COMMIT, the most common meaning is to ‘pledge/dedicate/devote’, with 58% of the occurrences in a random sample of 300, and features predominantly in positive and neutral environments, e.g. (1).

In contrast, the other prevalent meaning, synonymous with ‘perpetrate’, constitute 38% of the instances, but features exclusively in negative contexts, e.g. (2), which also holds true for the last and less common meaning of ‘imprison/place in custody’, e.g. (3). This chasm is also evident in the semantic preferences of the latter meanings, primarily connected to law and order, differing from the first meaning.

(1) The Government is also committed to new road safety measures

(2) Murder, assault, and theft are committed by working-class men

(3) He was eventually arrested and committed to the Crown Court

The previously ascribed prosody can therefore be attributed to these last two meanings of COMMIT, showing the importance of taking polysemy into account. These units now differ to such an extent that it may even be possible to regard them as semantically unrelated, despite their historical relatedness.

**References**

Bublitz, W. 1996. Semantic prosody and cohesive company: somewhat predictable. Leuvense Bijdragen: Tijdschrift voor Germaanse Filologie 85(1-2): 1-32

Oxford English Dictionary Online. 2022. Oxford University Press. commit, v. : Oxford English Dictionary (uio.no) [accessed 30 December 2022]

Partington, A. 1998. Patterns and Meanings. Using Corpora for English Language Research and Teaching. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Sinclair, J. 1996. “The search for units of meaning”. Reprinted in J. Sinclair & R. Carter (eds.), Trust the Text (2004), pp. 24-48. London: Routledge.

Stewart, D. 2010. Semantic Prosody. A Critical Evaluation. New York: Routledge.

Stubbs, M. 2002. Words and Phrases. Corpus Studies of Lexical Semantics. Oxford: Blackwell Publishing

**Keywords:** Semantic prosody, Discourse prosody, Collocation, Colligation, Semantic preference, BNC, British National Corpus

[3438] ***Massive Open Online Courses (MOOCs): Lexical Demands and Academic Vocabulary.***

Chen-Yu Liu (National Central University).

**Abstract.** Massive Open Online Courses (MOOCs) are among one of the most popular online learning resources either as supplements in the traditional classroom or as self-directed learning materials. They are considered helpful in terms of the learning of specialized knowledge as well as academic English. Yet, their lexical profile has remained underexplored. As vocabulary knowledge plays an essential role in facilitating comprehension, this study thus aims to examine the lexical demands and the coverage of academic vocabulary in MOOC lectures to provide pedagogical suggestions in terms of the use of such resource. Drawing on a 10-million-word corpus comprising MOOC lectures of four disciplinary areas (Engineering, Humanities and Arts, Science and Math, and Social Sciences), this study examines (a) the lexical demands of MOOC lectures, (b) the coverage of the Academic Spoken Word List (ASWL) in these lectures, and (c) the disciplinary variation in terms of lexical demands and academic vocabulary coverage among these lectures.

Results showed that together with proper nouns and marginal words, the most frequent 4,000 word families covered 96.36%, and the most frequent 8,000 word families covered 98.23% of the MOOC lectures. The ASWL had high coverage over the MOOC lectures (89.88%), suggesting that studying the ASWL can provide great help for students to use such resource. The analysis also revealed disciplinary variations in terms of both lexical demands and the ASWL representations among the MOOC lectures, with the Science and Math lectures being the most lexically demanding among the four disciplinary areas, while Social Sciences lectures having the highest ASWL coverage among all. This study expands the current understanding of MOOCs from a lexical perspective, and further provides pedagogical suggestions of using MOOC lectures to students with different vocabulary knowledge and disciplinary backgrounds.

**Keywords:** Massive Online Open Courses (MOOCs), Lexical demand, Academic spoken vocabulary, Disciplinary variation

[3451] ***What does it do, to who? Defining Identity in media representations of illness.***

Ewan Hannaford (University of Glasgow).

**Abstract.** Research demonstrates media coverage is a key source of health information for the public and substantially impacts public health attitudes, stigmas, and stereotypes (e.g. Wahl, 1992; Seale, 2003; Klin & Lemish, 2008; Loch & Rössler, 2017; Ross et al., 2019). The Common-Sense Model (CSM) theorises individuals draw on a wide range of informational sources to cognitively represent conditions across several key dimensions: ‘Cultural, institutional, social and personal factors influence each and every one of the variables involved in representation of health threats ... The cultural framework is set in bricks and mortar, in language and social relationships, as well as in culture-wide myths of cause and control.’ (Leventhal et al., 1997: 37–38). However, despite discursive practices being identified as integral to the construction of illness representations, the mechanisms by which these are manifested linguistically are unexplored in prior research; this talk presents results from completed doctoral research that began to correct this absence.

One of the key dimensions along which illnesses are thought to be conceptualised is Identity, which relates to the characteristics and labels associated with illnesses (Leventhal et al., 1997; Leventhal, Philips & Burns, 2016), including broader categorisations of illness (e.g. mental and physical). This talk examines how illness Identity may be constructed through media discourse by discussing findings of the comparative analysis of a 35 million-word corpus of nine illness discourses in the UK and US press. Consequently, it proposes a range of lexical, syntactic, and structural features that may establish differing public conceptualisations of Identity and how these may be best identified in analysis of media discourse, exploring how critical discourse analysis can be most effectively married with corpus linguistics in the spheres of health and media discourse.

**References**

Klin, A. & Lemish, D. 2008. Mental disorders stigma in the media: Review of studies on production, content, and influences. Journal of Health Communication 13(5). 434–449.

Leventhal, H., Benyamini, Y., Brownlee, S., Diefenbach, M., Leventhal, E.A., Patrick- Miller, L., & Robitaille, C. 1997. Illness representations: Theoretical foundations. In J. Weinman & K. Petrie (eds.), Perceptions of Health and Illness, 19–45. London: Harwood Publishers.

Leventhal, H., Phillips, L.A., & Burns, E. 2016. The Common-Sense Model of self-regulation (CSM): A dynamic framework for understanding illness self-management. Journal of Behavioral Medicine 39(6). 935–946.

Loch, A. & Rössler, W. 2017. Who is contributing? In W. Gaebel, W. Rössler, & N. Sartorius (eds.), The Stigma of Mental Illness: End of the Story?, 111–122. Cham: Springer International.

Ross, A.M., Morgan, A.J., Jorm, A.F., & Reavley, N.J. 2019. A systematic review of the impact of media reports of severe mental illness on stigma and discrimination, and interventions that aim to mitigate any adverse impact. Social Psychiatry and Psychiatric Epidemiology 54(1). 11–31.

Seale, C. 2003. Health and media: An overview. Sociology of Health and Illness 25(6). 513– 531.

Wahl, O.F. 1992. Mass media images of mental illness: A review of the literature. Journal of Community Psychology 20(4). 343–352.

**Keywords:** corpus linguistics, discourse analysis, health discourse, media discourse, illness representations, identity

[3481] ***‘Cancer pounces, snatches, latches on’ Mapping metaphor in patients’ narratives from a multilingual perspective: a corpus-based study.***

Jennifer Moreno (University of Zaragoza).

**Abstract.** Throughout the last decade, new communication tools have been implemented in healthcare as part of the efforts made to promote patient-centred care. This is known as narrative medicine, and its primary objective is to improve medical practice through the use of storytelling. On the one hand, storytelling provides us with the opportunity to view and experience events through the eyes of someone else, allowing us to understand a certain event as a meaningful experience.

On the other hand, as writers, storytelling offers us the opportunity to freely contemplate and express our ideas.

A growing popularity of patients’ narratives in recent years can be attributed to the advent of the Internet, which has made interaction and information exchange easier and has made this issue even more relevant. In this context, social media and patient-oriented websites have helped to raise awareness of the importance of telling stories, especially for patients. As a result, in the past few years, we have observed a significant increase in patients who are willing to share their illness experiences, particularly when they have chronic or serious illnesses, as they are suffering from cancer or other serious diseases.

In relation to this issue, it is important to emphasise that metaphor is often used as a means of describing subjective experiences that are difficult to describe. In view of the fact that cancer is one of these experiences, observing how metaphor is used in different languages to describe illness experiences is beneficial when it comes to understanding how individuals with different socio-cultural and linguistic backgrounds understand these experiences. This paper presents the results of a corpus-based study that assessed how metaphors are used by patients when discussing illness in Italian, Spanish, and English narratives available online.

**Keywords:** narrative medicine, patients’ narratives, storytelling, illness experience, metaphor, corpus-based study

[3592] ***Creating a Frequency Dictionary of Current British English: Methodological Considerations.***

Vaclav Brezina (Lancaster University) and Dana Gablasova (Lancaster University).

**Abstract.** It is relatively easy to produce a corpus-based wordlist; it is much more difficult to make a wordlist usable for pedagogical and research purposes (Brezina & Gablasova, 2017). Creating a frequency dictionary (Brezina & Gablasova, forthcoming), brings yet another level of complexity.

In this paper, we reflect on a series of assumptions underlying a construct of a frequency dictionary of English as well as on a series of methodological decisions the compilers of such a dictionary are faced with. In particular, we will focus on the following five areas: i) units of analysis, ii) order of items, iii) dispersion measures, iv) inclusion of proper nouns and v) multi-word expressions.

The frequency dictionary is based on the data from the British National Corpus 2014 (BNC2014), a 100-million-word representative sample of current British English developed at Lancaster University (Brezina et al. 2021; Love et al. 2017). The BNC2014 was constructed as a comparable counterpart to the original British National Corpus (BNC1994), which was compiled in the early 1990s. The BNC2014 is a balanced sample of all major genres/registers.

The dictionary necessarily reflects the composition of the source data: what texts have been included and in what proportions. Understanding the sources and their proportions is therefore important for proper interpretation of the final product. In addition, we need to bear in mind that an automated procedure was used to identify headwords (lemmatisation) and word class membership (pos-tagging) of lexical items. This brings a certain level of inaccuracy and variation. While the estimated overall precision rate is 96% (Leech and Smith, 2000) we note cases where manual correction was necessary in order to maintain high accuracy level of the material presented.

**Keywords:** vocabulary, wordlists, methodology

[3593] ***Things That Make Us Go Ooh… (Why We Keep Rolling Up to the Circus).***

Katharine Kavanagh (Cardiff University).

**Abstract.** The early 21st Century UK has a thriving and expanding circus arts sector. What do audience members value about their experience now, when so much public representation of circus is based on mid-20th Century models? This research investigates an evaluative corpus of circus publicity and circus audience experience texts to identify such values. Moreover, it examines gaps between audience member priorities and those publicly presented in promotional discourse.

The method combines corpus linguistic approaches with an analytical framework borrowed from Benitez-Castro and Hidalgo-Tenorio's (2019) adaptation of Martin and White’s (2005) APPRAISAL system, allowing recurring targets of evaluation to be reframed as audience values. The corpus comprises 100,272 tokens of downsampled review texts published between 1 February 2019 and 31 January 2020, 38,958 tokens of corresponding publicity material texts, and 24,177 tokens of audience interview text collected at the Glastonbury Festival Circus Big Top venue in 2022 and transcribed following Collins and Hardie’s (2022) recommendations.

Key semantic domain analysis was conducted using WMatrix (Rayson, 2008), with resultant domains divided into formal, ideational and evaluative categories. Lexical items from evaluative categories were used as concordance search terms within Sketch Engine (Kilgariff et al, 2014), and the results were extracted and combined into a new analysis corpus that was manually annotated in the UAM CorpusTool (O’Donnell, 2008).

Preliminary findings indicate that qualities of surprise, captivation, and variety are highly valued elements of circus audience experience. However, review texts are ideationally aligned more with publicity blurbs than with direct audience member evaluations. These public texts reflect and normalise the values of the dominant performing arts industry, emphasising elements of experience that are accorded status within the legitimised theatre sphere rather than those pertinent to the historically popular form of circus. In this way, a division between elite and popular forms of performance appears to be reified, further evidencing Baker et al’s (2008) ‘useful methodological synergy’ of corpus linguistics with CDA.

**Keywords:** evaluation, performing arts, cda, key semantic domains, spoken corpora, promotional texts

[3647] ***“Tell me you want it, sissy”: Characterisation, narrative structure and the troubling of agency in sissy porn.***

Alon Lischinsky (Oxford Brookes University) and Kat Gupta (Roehampton University).

**Abstract.** There is considerable scholarship on the ways in which effeminacy has been discursively stigmatised in modern Western culture (eg, McInnes & Davies, 2008; Rottnek, 1999; Robinson & Davies, 2007). Analysis of both mediated representations and face-to-face interactions shows that those perceived as ‘sissies’ tend to attract intensely negative responses of censure and revulsion (Corbett, 2009: 362; Kuhl & Martino, 2018: 31; Thorne, 1993: 116).

But,though hegemonic, this negative conceptualisation of male femininity is not universal. A small but growing body of research has begun to explore how effeminacy can be alternately conceptualised as “a source of desire, power, and/or transcendence of rigid, heteronormative sex/gender roles” (Thomas, 2017:5; cf. Goldberg, 2021; Messias, 2011; Vytniorgu, 2022).

In this paper, we build on this tradition by exploring the characterisation and evaluation of sissy characters in erotic writing, a genre in which issues of gender play, attraction and fantasy occupy a central role.

Using a corpus of approximately 1.4 billion word-tokens collected from Literotica.com (2016), one of the oldest, largest and most widely-read erotic fiction repositories online, we use verbal, nominal and adjectival collocates to characterise the semantic profile of the term ‘sissy’ and compare it to a range of other common terms for gender-nonconforming characters with which it may co-occur, such as ‘queer’, ‘tranny’ or ‘boi’.

Our analysis shows that — despite the occasional interchangeable use — sissy is a semantic outlier among these terms, associated with terminology denoting Dominance/submission play and nonmonogamous relationship dynamics, often with a strong non-consensual element.

Drawing on Bamberg’s (2010) concept of “dilemmatic spaces” through which selfhood is navigated, we discuss how sissy characters' ambiguous agency about their own identity echoes the compulsory nature of hegemonic masculinity.

**Keywords:** corpus stylistics, corpus-assisted discourse studies, language and sexuality, gender studies, porn studies, effeminacy

[3763]***Two approaches to collocation networks: GraphColl and LLN.***

Hanna Schmueck (Lancaster University) and Vaclav Brezina (Lancaster University).

**Abstract.** This paper compares two approaches to the creation and analysis of collocation networks: the GraphColl function currently available through #LancsBox and #LancsBox X (Brezina et al., 2020) and a Python and Cytoscape-based (Shannon et al., 2003) custom pipeline for visualisation and graph theoretical analysis of corpus-wide linguistic networks (LLN). Collocation networks are used to visualise discourse relationships, "aboutness" of a text or discourse, semantic relations (Pecina, 2010; Xiao & McEnery, 2006; Brezina, 2016; Baker, 2016; Brezina et al., 2015) as well as lexicogrammatical features (McEnery & Brezina, 2019).

GraphColl is an accessible tool to view collocational relationships localised to a particular area of discourse; it allows selecting a particular set of search terms, association measure parameters, window spans and thresholds, and retains information on the type, lemma and POS level before plotting collocation networks. LLN, on the other hand, is the result of a methodological triangulation (Noble & Heale, 2019; Tucker & Ernestus, 2016) and integration synthesis (Wyatt, 2022) of concepts from corpus linguistics and graph theory. A core difference is that LLN yields graph theoretical parameters that describe all collocations found in the entire corpus thus producing much larger networks. It also incorporates a graph theoretical exploration e.g. in order to compare different corpora via their collocational density, the average distance between their collocates, etc. Beyond these global descriptions, LLN also allows for clustering to explore different ‘layers of discourse’.

This paper will provide an in-depth demonstration of how the collocation networks generated by the two tools differ. The case study will focus on the topic of substance abuse across discourses represented in the BNC2014 (Brezina et al., 2021; Love et al., 2017). The implications for corpus-based discourse analysis will be discussed.

**References**

Baker, P. (2016). The shapes of collocation. International Journal of Corpus Linguistics, 21(2), 139–164. <https://doi.org/10.1075/ijcl.21.2.01bak>

Brezina, V. (2016). Collocation Networks: Exploring Associations in Discourse. In P. Baker & J. Egbert (Eds.), Routledge Advances in Corpus Linguistics: Vol. 17. Triangulating methodological approaches in corpus-linguistic research (pp. 90–107). Routledge.

Brezina, V., Hawtin, A., & McEnery, T. (2021). The Written British National Corpus 2014 – design and comparability. Text & Talk, 41(5-6), 595–615. <https://doi.org/10.1515/text-2020-0052>

Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. International Journal of Corpus Linguistics, 20(2), 139–173. <https://doi.org/10.1075/ijcl.20.2.01bre>

Brezina, V., Weill-Tessier, P., & McEnery, A. (2020). #LancsBox v. 5.1 [Computer software]. <http://corpora.lancs.ac.uk/lancsbox>.

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). Compiling and analysing the Spoken British National Corpus 2014. International Journal of Corpus Linguistics, 22(3), 319–344. <https://doi.org/10.1075/ijcl.22.3.02lov>

McEnery, T., & Brezina, V. (2019). Collocations and colligations: Visualizing lexicogrammar. In B. Busse & R. Moehlig-Falke (Eds.), Topics in English linguistics: Vol. 104. Patterns in language and linguistics: New perspectives on a ubiquitous concept (pp. 97–124). De Gruyter Mouton. <https://doi.org/10.1515/9783110596656-005>

Noble, H., & Heale, R. (2019). Triangulation in research, with examples. Evidence-Based Nursing, 22(3), 67–68. <https://doi.org/10.1136/ebnurs-2019-103145>

Pecina, P. (2010). Lexical association measures and collocation extraction. Language Resources & Evaluation, 44, 137–158. <https://doi.org/10.1007/s10579-009-9101-4>

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>

Tucker, B. V., & Ernestus, M. (2016). New Questions for the Next Decade. The Mental Lexicon, 11(3), 375–400. <https://doi.org/10.1075/ml.11.3.03tuc>

Wyatt, S. (2022). Interdisciplinarity: Models and Values for Digital Humanism. In H. Werthner, E. Prem, E. A. Lee, & C. Ghezzi (Eds.), Perspectives on Digital Humanism (pp. 329–334). Springer International Publishing.

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>

**Keywords:** collocation network, collocations, graph theory, network, substance abuse, BNC2014

[3820] ***Why do function word frequencies vary across individuals? Evidence in favour of The Statistical Approximation Hypothesis.***

Andrea Nini (University of Manchester).

**Abstract.** A modern unsolved mystery in applied linguistics is that the frequency of function words can be effectively used to identify the author of a text. Computational methods to identify authors were developed by computer scientists or statisticians who chose to focus on function words because of their high frequency and topic obliviousness (Mosteller & Wallace 1963). However, linguistically speaking, it does not make sense to say that authors differ in their function word frequency because of personal preference. Another reason why this explanation does not make sense is that, as corpus linguistic evidence has always suggested (Sinclair 2004), the basic unit of language is not the word but something larger and more similar to a phrase, a claim compatible with cognitive linguistic and cognitive psychology evidence on language processing (Christiansen & Chater 2016; Langacker 1987).

In this paper I introduce a formal theory of linguistic individuality based on cognitive linguistic principles that could offer an explanation in the form of the Statistical Approximation Hypothesis. This hypothesis says that the frequency of function words reflects the distribution of function words in our repository of linguistic units, which, according to usage-based frameworks, should be unique because of the idiosyncratic process of entrenchment (Schmid 2015). If this hypothesis is correct, then authorship analysis methods based on this principle should outperform or perform equally as methods based on function word frequencies. Preliminary evidence based on analysis of existing

benchmark corpora for authorship analysis indeed shows that this is the case. New methods simply based on the presence or absence of n-grams perform as well as a sophisticated machine learning or deep learning algorithms that consider frequency of function words. This result is interpreted as indirect evidence in favour of the proposed hypothesis.

**References**

Christiansen, Morten H. & Nick Chater. 2016. The Now-or-Never bottleneck: A fundamental constraint on language. Behavioral and Brain Sciences. Cambridge University Press 39. e62. <https://doi.org/10.1017/S0140525X1500031X>.

Langacker, Ronald W. 1987. Foundations of cognitive grammar. Stanford, CA: Stanford University Press.

Mosteller, Frederick & David L Wallace. 1963. Inference in an Authorship Problem. Journal of the American Statistical Association 58. 275–309. <https://doi.org/10.2307/2283270>.

Schmid, Hans-Jörg. 2015. A blueprint of the Entrenchment-and- Conventionalization Model. Yearbook of the German Cognitive Linguistics Association 3(1). 3–25. <https://doi.org/10.1515/gcla-2015-0002>.

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

**Keywords:** authorship analysis, forensic linguistics, idiolect, linguistic individuality, function words, n-grams

[3940] ***Women in Early Modern Manuscript News: Harnessing a Lost Corpus with Bootstrapping and Bayesian Statistics.***

Gabor Mihaly Toth (University College, Cork).

**Abstract.** In early modern Europe information circulated in the form of regular semi-public manuscript newsletters. However, most of the once existing news sheets did not survive. Those news sheets that survived are dispersed in different collections of the world. These collections are just random and incomplete samples from the once existing corpus of handwritten news. How can we unlock and study the lost early modern corpus of handwritten news? How can we study its gender dimension (for instance, presence of women)?

This paper will address these questions by presenting a completed computational study of a large digitised corpus of transcribed handwritten news sheets in Italian from the Medici Archive. The corpus consists of 1200 news sheets covering a period of 200 years and compiled in the major urban centres of early modern Europe.

First, the paper will offer a concise introduction into handwritten news sheets. Second, it will demonstrate how random sampling with replacement (aka bootstrapping) helped harness the randomness of news sheets preserved in the Medici Archive. Third, it will discuss the problem of how to get from the evidence collected with random sampling of news sheets to the lost corpus of news sheets circulating in the early modern period. Precisely, the paper will focus on the presence of women in the lost corpus. The problem will be framed in terms of Bayesian statistics as follows: women’s presence in the lost corpus is the Prior; the result of bootstrapping is the Evidence. The paper will show how Bayesian formula, Bayesian updating, and the beta distribution as conjugate prior, enabled the development of a plausible model representing women’s presence in the lost corpus of early modern news sheets.

In the conclusion, I will reflect on the broader problem of how random sampling and Bayesian inference can help reconstruct lost corpora.

**Keywords:** Historical corpora, Bayesian statistics, Random sampling, Loss and incompleteness, Gender history

[3963] ***Exploring Saudi Novels through the Lens of Corpus Linguistics.***

Tareq Alfraidi (Islamic University of Madinah).

**Abstract.**

Introduction:   
Studying the language of literature through a corpus linguistic perspective has recently been recognized as a valuable approach to stylistics studies. It provides the research community with fresh results that were difficult to gain when relying on intuition (Stubbs, 2005; Fischer-Starcke, 2010; Mahlberg et al. 2016; Kubis, 2021). This field of research is normally referred to as 'Corpus Stylistics' (Mahlberg et al. 2015, Alfraidi et al. 2022).

Although a considerable number of studies have applied the corpus stylistics approach to the analysis of English novels and other European languages, it has been neglected in the study of Saudi novels, where the Modern Standard Arabic variety is mainly used. The dominant approaches used are primarily ‘traditional’, where intuition and subjective judgment are largely employed (Alfraidi et al. 2022). Evidently, this indicates a noticeable gap in this area. Therefore, this paper aims to offer a pioneering analysis of the language used in Saudi novels via the lens of corpus linguistics. It demonstrates how corpus techniques are advantageous in the investigation of the style of Saudi novels. It attempts to answer, in the main, questions related to the frequency of words, keywords, their textual patterns and their meanings.

Data and methods:

The data used in this study is drawn on the Saudi Novel Corpus (SNCorpus). It is regarded the first corpus built specifically to reflect the language usage of Saudi novels (Alfraidi et al. 2022). Currently, the corpus contains 53 novels that comprise collectively of around 3.1 million words (ibid.). The analysis was conducted on a segmented version of the corpus, and the free Arabic corpus processing tool Ghawwas (Al-Thubaity et al., 2014) was used to generate the results.

The analysis and main findings:

As a first step in the analysis, we extracted a list of the top 100 content words to obtain a general picture of the most frequent lexical items used by Saudi novelists. We noticed that around 60% of these words are nouns. In addition, we observed that the noun nafs (self) is at the top of the list.

A closer look at the concordance line of this word revealed the common textual patterns associated with its use. For example, this word functions as a reflexive marker, emphasizer, or synonym for 'soul', 'spiritual', or human being(s), yet the first function is the dominant one. We also detected that the word Allah (the God) is ranked second among the top words. This high frequency is due to the pervasiveness of Islamic culture in the narratives of Saudi novels. Evidence of this interpretation of the results is observed through our examination of the concordance lines that showed this word appearing frequently in formulaic Islamic phrases (e.g. insha allah (God willing), alhmadu lillah (praise to be God) uttered commonly by the novel's characters. This establishes a parallelism between written fiction and spoken mode, which makes what is narrated fictionally closer to reality.

To gain a better picture of the distinctive features of the language used in Saudi novels, we generated a list of the keywords and analyzed them. This was performed by comparing the SNCorpus wordlist with the Open Source Arabic Corpora (OSAC) wordlist. As a result, some noteworthy findings have been gained regarding keywords, key grammatical categories, and key semantic fields. For example, the possessive pronoun ي (my) appears at the top of the list, accounting for the fact that the expression of 'possession' is highly dominant and reflects the style of the Saudi novel genre. In addition, the words related to the 'human body' are unexpectedly frequent, while the noun وجه (face) scores the highest degree of keyness. Considering the function words, we found that interrogative particles demonstrate higher signification frequencies in the corpus compared to other function words.

Our hope is that these results offer fresh insights into the features of the language used in the Saudi novel genre and reflect the stylistic characteristics usually absent in other Arabic genres. More quantitative and qualitative results regarding the most frequent words and keywords and their textual patterns will be reported in the paper.

**References**

Al-Thubaity, A., Khan, M., Al-Mazrua, M., & Almoussa, M. (2014b). KACST Arabic Corpora Processing Tool ‘‘Khawas’’ [Computer Software]. Retrieved July 2, 2016, from <https://sourceforge.net/projects/ghawwasv4/>.

Fischer-Starcke, B. Corpus Linguistics in Literary Analysis: Jane Austen and Her Contemporaries; Bloomsbury Publishing: London, UK, 2010.

Kubis, M. Quantitative analysis of character networks in Polish 19th-and 20th-century novels. Digit. Scholarsh. Humanit. 2021, 36, ii175–ii181.

Mahlberg, M.; Biber, D.; Reppen, R. Literary style and literary texts. In The Cambridge Handbook of English Corpus Linguistics; Cambridge University Press: Cambridge, UK, 2015; pp. 346–361

Mahlberg, M.; Stockwell, P.; Joode, J.d.; Smith, C.; O’Donnell, M.B. CLiC Dickens: Novel uses of concordances for the integration of corpus stylistics and cognitive poetics. Corpora 2016, 11, 433–463.

Saad, M.K.; Ashour, W.M. Osac: Open source arabic corpora. In Proceedings of the 6th ArchEng International Symposiums (EEECS), Opatija, Croatia, 12–15 December 2010; Volume 10.

Stubbs, M. Conrad in the computer: Examples of quantitative stylistic methods. Lang. Lit. 2005, 14, 5–24.

**Keywords:** Corpus stylistics, Saudi novels, Arabic, keywords, textual patterns

[4188] ***Variation in Response Structure in a Corpus of Online Chat.***

Jack Grieve (University of Birmingham), Sara Bartl (University of Birmingham), Solly Elstein (University of Birmingham), Kira Ogburn (University of Birmingham) and Dana Roemling (University of Birmingham).

**Abstract.** Although conversation has long been the subject of careful and focused analysis in conversation analysis, discourse analysis, and interactional sociolinguistics, we lack a broader empirical understanding of variation in the overall organisation of conversations.

In this presentation, we introduce a general framework for representing the structure of a conversation, focusing on what we call response structure: we link every turn in the conversation back to the previous turn or turns from an interlocutor, if any, to which it was produced in direct response. For example, an answer is linked back to a question (as all second pair parts are linked back to first pair parts in adjacency pairs as analysed in conversation analysis). We then define the overall structure of a conversation as the set of all response links between turns. This allows for conversations to be represented as a series of one or more response-link trees, like a syntactic analysis of a sentence, except that turns instead of words act as leaf nodes.

Based on this system, we then describe our manual coding of 200 conversations drawn from the Ubuntu Dialogue Corpus – a large corpus of Internet Relay Chats where novice users seek technical help with the Ubuntu operating system in dyadic online text-based interactions with expert users. We present numerous examples of parsed conversations from the corpus to illustrate the application of our framework.

Finally, we introduce various metrics for comparing the structure of the response-link trees (e.g. average response link length), and we use this information to identify dimensions of variation in the structure of the 200 Ubuntu Dialogue Corpus chats. We find that conversations differ broadly in terms of the linearity of their response links, and we consider the functional correlates of this variation in overall conversational structure.

**Keywords:** Chat, Computer Mediated Communication, Conversation, Conversation Analysis, Discourse Analysis, Interaction

[4314] ***Unveiling the Function(s) and Positions of English Discourse marker Now and its Persian Counterpart Hâlâ: A Contrastive Corpus-based Inquiry.***

Zahra Ghane (Heinrich Heine University) and Mehrdad Vasheghani Farahani (Leipzig University).

**Abstract.** Depending on their positions, discourse markers can receive various functions in spoken and written modes. Given this fact, the present study set out to investigate the functions of the English discourse marker now and its Persian Counterpart hâlâ in initial, medial, and final positions in English and Persian languages. In more detail, it seeks to answer these questions: 1. What are

the functions of now in English and its counterpart hâlâ in Persian in three initial, medial and final positions? 2. Is there any relation between the position of the functions of the two discourse markers? And if yes, why and how? To this end, two fine-grained corpora of the British National Corpus (BNC) (100 million words) and the Bijankhan Corpus ( 86,560,954 words) were chosen to derive the relevant data in English and Persian. The application of Kaltenböck, et. al. (2011) and Heine (2013) theoretical frameworks in scrutinizing the concordance lines in these two corpora indicated that the function of these two discourse markers influences the position(s) they can appear in. These markers come to highlight a kind of transition the addressee has faced or is to face. These transitions can canonically occur in the (sub)topic the speaker is talking about, but is extendable to some other aspects of discourse as speech act, implicature, mood, turn and term. The results of this study can have practical implications for researchers in text analysis, corpus linguistics and contrastive analysis.

**Keywords:** Discourse Markers, Function, Position, Transition, English and Persian Languages.

[4374] ***Gayborhoods, Queeristan, and bi-nosaurs – the conceptual blending of queer terms in The TV Corpus.***

Charlotte Rosenow (RWTH Aachen University).

**Abstract.** Due to the carefully planned, practiced, and edited nature of scripted television, its use of language and, as such, “TV dialogue” (Bednarek 2018, 7) can be assumed to allow us to understand “how language used in television texts connects to a world outside the text.” (Marshall and Werndly 2002, 94).

Similarly, the influence that behaviours witnessed on TV may have on the opinions of a viewer has been well documented, especially with regards to the representation of marginalized groups (e.g. Pugh 2018). Meanwhile, the use of blending as a word formation strategy reportedly is extremely productive, and conceptual blending (Fauconnier and Turner 2008) is seen as particularly well-suited to analysing new blends (e.g. Schmid 2016).

This research investigates terms of queer in-group identity in scripted North American TV shows in The TV Corpus (Davies 2019) as part of an ongoing PhD research project. This paper in particular focuses on blending as a word formation strategy, discussing both its prevalence within extracted datasets from a quantitative perspective, as well as qualitatively assessing these conceptual blends in order to better understand how they are used to represent the queer community on scripted TV.

**References**

Bednarek, Monika. 2018. Language and Television Series: A Linguistic Approach to TV Dialogue. Cambridge University Press.

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

Fauconnier, Gilles, and Mark Turner. 2008. The way we think: Conceptual blending and the mind's hidden complexities. New York: Basic Books.

Marshall, Jill, and Angela Werndly. 2002. The Language of Television. Intertext. London ; New York: Routledge.

Pugh, Tison. 2018. The Queer Fantasies of the American Family Sitcom. New Brunswick, NJ: Rutgers University Press.

Schmid, Hans-Jörg. 2016. English Morphology and Word-Formation. Vol. 25. Berlin: Ludwig-Maximilians-Universität München. <https://epub.ub.uni-muenchen.de/42485/>.

**Keywords:** queer studies, language use on scripted television, conceptual blending

[4390] ***"I'm so OCD lol": A corpus study of OCD as an adjective.***

Jordan Batchelor (Georgia State University) and Heewon Lee (Georgia State University).

**Abstract.** Obsessive compulsive disorder (OCD) is characterized by the presence of recurrent thoughts or images that intrude a person’s psyche, resulting in repetitive behaviors that mitigate the distress caused by such thoughts or images (APA, 2013). While studied widely from a psychiatric perspective, one issue that has gained attention in popular discourse is the use of OCD as an adjective (e.g., I’m so OCD), which is thought to trivialize the symptoms of the disorder (NAMI, 2015). However, little research has examined this issue empirically, specifically how frequently it occurs and in what semantic domains it occurs. To contribute toward answering these questions, we collected a corpus of 88,000 recent social media comments including the word “OCD” and identified all uses of OCD modified by a degree adverb. We then applied n-gram analysis and collocational analysis (Mutual Information 2) to better understand the contexts in which the phrase was used most. We found 1,575 comments including the phrase degree adverb + OCD, most of which (~75%) used the phrase genuinely to describe a feature of themselves or someone else, while a smaller number (~25%) used the phrase ironically, deriding its genuine use by others. While ironic comments frequently mentioned the phrase so OCD, genuine comments used a wider variety of (mostly) intensifying adverbs (e.g., very, too, and super). N-gram analysis revealed many frequent 5-grams of the construction I’m + degree adverb + OCD + about. Collocations of these 5-grams related to cleanliness (e.g., hygiene, cleanliness), general minutia (e.g., details, numbers), and certain hobbies (e.g., shoes, cars). Our analyses support the popular belief that the adjectival use of OCD tends to occur in the context of cleanliness, organization, and hobbies, and that such comments de-medicalize the symptoms of OCD.

**Keywords:** health communication, corpus-based discourse analysis, Obsessive compulsive disorder (OCD)

[4427] ***Artificial intelligence: a corpus-assisted discourse analysis across multiple stakeholders.***

Jekaterina Nikitina (University of Milan).

**Abstract.** Starting from 2010s, massive use of technology and the increased processing power of computers fuelled the boom of AI systems. Against the multifaceted and pervasive nature of AI, it is particularly important to look how AI-related meanings are communicated in specialised contexts (Garzone 2020), across different stakeholders and whether different meaning facets, or even ideologies, transpire. This study investigates how AI is linguistically represented through terms, multi-word terms and collocations in legal documents produced by international organisations (the UN, OECD, the Council of Europe, EU institutions, and others) and by corporations dealing with AI (Google, Microsoft, Apple, etc) between 2020 and 2022, focussing on archetypal representations as well as convergent and divergent strategies. The corpus consists of reports, studies, roadmaps, as well as principles, guidelines and websites.

The analysis assesses the qualitative selection of AI-related representations (what terms and collocations are chosen as connected to AI) as well as the quantitative dimension (how prominent or dispersed are some representations). The study adopts Corpus-Assisted Discourse Analytical approach (Partington 2004; Baker et al. 2008; Nikitina 2020), exploring the synergy of Corpus Linguistics (Stubbs 1994; McEnery and Brezina 2022) with the (critical) discourse analytical perspective to outline the interplay between discourse, ideology and power (Fairclough 2014; van Dijk 1998).

Preliminary findings indicate that most stakeholders – when defining and describing AI, its possible implications and applications – converge on the importance of transparency, privacy, human rights, responsibility and trust, yet assign different weight to them. As some institutions announced their plan to create a universal binding legal instrument defining AI, these findings may contribute to a more balanced concept representation – or definition – in the near future.

**References**

Baker P., Gabrielatos C., Khosravinik M., Krzyzanowski M., McEnery T., and Wodak R. 2008, “A Useful Methodological Synergy? Combining Critical Discourse Analysis and Corpus Linguistics to Examine Discourses of Refugees and Asylum Seekers in the UK Press”, Discourse and Society 19 (3): 273-306.

Fairclough N. [1989] 2014, Language and Power, London: Longman.

Garzone, G. 2020. Specialized Communication and Popularization in English. Roma: Carocci Editore.

McEnery, T. and Brezina, V. 2022. The Fundamental Principles of Corpus Linguistics. Cambridge: Cambridge University Press.

Nikitina, J. 2020. “Representation of Gene‐Editing in British and Italian Newspapers: A Cross‐Linguistic Corpus‐Assisted Discourse Study.” Lingue e Linguaggi 34: 51–75.

Partington A. 2004, “Introduction: Corpora and Discourse, A Most Congruous Beast”, in Partington A., Morley J. and Haarman L. (eds), Corpora and Discourse, Peter Lang, Frankfurt am Main, 11-20.

Stubbs M. 1994, “Grammar, Text, and Ideology: Computer-assisted Methods in the Linguistics of Representation”, Applied Linguistics 15 (2): 201-223.

van Dijk T. 1998, Ideology: A Multi-disciplinary Approach, Sage, Thousand Oaks (CA).

**Keywords:** corpus-assisted discourse analysis, Artificial Intelligence, discourse of international organisations international corporations

[4525] ***Build your own word family: An adaptable vocabulary profiler for French, German and Spanish.***

Natalie Finlayson (University of York), Emma Marsden (University of York) and Laurence Anthony (Waseda University).

**Abstract.** The most appropriate unit of counting for evaluating lexical knowledge continues to be debated in corpus-based vocabulary studies.

Type counting is common in corpus research, but vocabulary researchers have traditionally used the word family unit, which relies on the assumption that learners can recognise links between derivationally related words. Recently, it has been suggested that the lemma or flemma might be the more appropriate unit across proficiency levels (e.g., McLean, 2018; Brown et al., 2020). However, studies seeking to provide evidence for the appropriacy of different types of lexical unit have yielded mixed results.

Now, the conversation is turning to the importance of varying the selection of units according to pedagogical purpose and learner variables (e.g., Cobb & Laufer, 2021; Webb, 2021; first proposed by Bauer & Nation, 1993). A bespoke approach is particularly relevant in the study of declined or highly inflected languages. Learners of such languages with limited knowledge of L2 grammar and a more typologically related L1 might find mastering a lemma a more challenging task than acquiring a set of regular, transparent affixes. Counting lemmas may simultaneously overestimate these learners’ ability to recognise inflected forms, and underrepresent their capacity to recognise derived forms.

In this talk, we present MultilingProfiler 3.0 (Finlayson et al., 2022), a vocabulary profiling tool designed to give corpus researchers and materials developers working in French, German, and Spanish more control over the lexical unit used to create profiles. Users can select which inflected and derived forms to include in the word definition, or choose a pre-made word list that uses a bespoke unit aligned with the grammar specifications of a particular curriculum or stage in a program of study. Results generated using these approaches are likely to better represent the lexical knowledge of the target learner group.

**References**

Bauer, L. & Nation, P. (1993). Word families. International Journal of Lexicography, 6, 253–279.

Cobb, T., & Laufer, B. (2021). The nuclear word family list: A list of the most frequent family members, including base and affixed words. Language Learning, 71, 834–871. <https://doi.org/10.1111/lang.12452>

Finlayson, N., Marsden, E., & Anthony, L. (2022). MultilingProfiler (Version 3) [Computer software]. University of York. Accessed 22 December 2022 at <https://www.multilingprofiler.net/>

McLean, S. (2018). Evidence for the adoption of the flemma as an appropriate word counting unit. Applied Linguistics, 39, 823–45. <https://doi.org/10.1093/APPLIN/AMW050>

Webb, S. (2021). A different perspective on the limitations of size and levels tests of written receptive vocabulary knowledge. Studies in Second Language Acquisition, 43, 454–461. <https://doi.org/10.1017/S0272263121000449>

**Keywords:** vocabulary, lexical profiling, foreign language education, materials development, language testing, French, German, Spanish

[4540] ***Discourse markers variants: A corpus study of Linguistics research article abstracts in Spanish.***

Hernán Robledo (Pontificia Universidad Católica de Valparaíso) and Carlos Aguilar (Pontificia Universidad Católica de Valparaíso).

**Abstract.** Formal invariability of discourse markers (DMs) is an inherited feature of the particle class (Pavón, 2003) and results from gradual processes of linguistic change, either considered lexicalisation or grammaticalisation (Garachana, 1998, 2008; Company, 2004; Brinton & Traugott, 2005; Degand & Simon-Vandenbergen, 2011; Onodera, 2011; Heine, 2013; Heine & Kaltenböck, 2021). In Spanish, most inventories of DMs include the more prototypical members of the category. Nevertheless, DMs with a lower degree of grammaticalisation can exhibit formal variation, and these cases tend to be less represented in the literature.

Considering this framework, this paper aims to describe the formal structure and discursive functions of ungrammaticalized forms of DMs used in research article abstracts (RAAs). This genre was chosen because the nature of RAAs remains problematic (Sánchez, 2009; Blanco, 2015; Piqué-Noguera and Camaño-Puig, 2015) and this work can contribute to their characterisation. Specifically, we worked with a corpus of RAAs published in Linguistics scientific journals in Spanish.

We analysed the corpus using a mixed method to identify and describe those DMs that show some degree of formal variability within a paradigm, i.e. they present an invariant and a variable part.

Among the results of the analysis we distinguish 13 groups of DMs showing formal variability within a paradigm: counterargumentative connective structures which are formal variations of a pesar de ‘in spite of’: ‘a pesar de + noun phrase’ and ‘a pesar de + relative clause’; purpose connective structures which are variations built with para ‘for’: para ello ‘for this’, ‘para + infinitive clause’ and ‘para + noun phrase’; commentators with structure en este contexto/sentido/línea ‘in this context/sense/line’; among others.

We also observed that some variants refer anaphorically to previous segments, often encapsulating and labelling previous information (López, 2014). These results contribute to a systematic characterisation of the formal variation present in Spanish DMs.

**Keywords:** Discourse markers, Connectives, Grammaticalisation, Research article abstacts

[4548] ***Language Learners’ Use of Conversational Persian: Insights from a Learner Corpus.***

Sepideh Daghbandan (University of Edinburgh).

**Abstract.** Conversational Persian is at its early stages of receiving attention in the field of Teaching Persian as a Second Language. However, research on the use of this register by learners of Persian remains scarce. Therefore, this study aims to explore the Conversational Persian used by language learners by employing a learner corpus methodology. To this end, a spoken learner corpus, the Learner of Persian Spoken Corpus (LoPSC), was compiled. LoPSC is the first spoken learner corpus of Persian. It is also the first learner corpus focusing specifically on Conversational Persian.

Data from LoPSC consists of approximately 40,000 words of transcribed audio recordings from conversations between advanced learners of Persian. Next, to gain a better understanding of the challenges that learners may encounter when using Conversational Persian, LoPSC was compared to a reference corpus, namely, the Conversational Persian Corpus. This corpus consists of 60,000 words of audio transcribed recordings from conversations between first language speakers of Persian. The results from the corpus-based analysis revealed that the most significant difference between the use of Conversational Persian by learners and first language speakers of Persian was

the word choice of these two groups of speakers. This difference in word choice was significantly reflected in the use of discourse markers. That is, the learners used significantly less discourse markers compared to their L1 speaker counterparts. In addition to the significant lower frequency, discourse markers appeared with less variations of forms in the LoPSC. Finally, the speakers of the two groups used different pragmatic functions for the same discourse markers.

This study has two main contributions. First, it provides empirical findings for a lesser studied variety, namely, the use of Conversational Persian by learners. Second, this study also provides insights into the challenges of compiling and analysing data from a spoken learner corpus of Persian.

**Keywords:** learner corpus, Persian Language, Discourse Markers

[4577] ***Thematic role predictors in German and Russian: a corpus study.***

Yevheniy Skyra (Max-Planck-Institute for Psycholinguistics), Rowena Garcia (University of Potsdam) and Evan Kidd (Australian National University).

**Abstract.** Languages use a variety of linguistic features to mark thematic roles (e.g., case, word order), and children must acquire them from their input. Even though two languages may have the same typological means for signifying thematic roles (e.g., case marking), there is inevitable language-specific variation in the reliability of cues in the input. The current research investigates this problem in two languages: German and Russian. We analyzed German and Russian child-adult conversations using conditional inference trees and random forests to investigate which linguistic information is available to children to assign thematic roles. Our results showed that word order is the most reliable cue to thematic role assignment in German, even though it is a case marking language. Coding and analysis of Russian is ongoing, but we predict that case will outrank word order. Our findings provide important insights into past experimental work.

**Keywords:** thematic role assignment, language acquisition, variation, cue reliability, recursive partitioning

[4633] ***Discourses emerging from anti-vaccine movements in Brazil on Twitter, during the COVID-19 pandemic.***

Mirella Whiteman (Pontifical Catholic University of São Paulo), Arianne Brogini (Pontifical Catholic University of São Paulo), Stephan Hughes (Pontifical Catholic University of São Paulo) and Paula Tavares Pinto (UNESP - São Paulo State University).

**Abstract.** Over the last decade, anti-vaccine movements have grown considerably with the aid of digital media. The return of previously controlled diseases such as measles and polio has become a major concern for health professionals around the world in general, and more recently, in Brazil. The coverage of the country's Inoculation Program has waned since 2015, taking a turn for the worse since the COVID-19 outbreak, due to denialist practices encouraged by the former federal government and its followers. As a result, vaccines and their effectiveness have taken center stage in the public conversation, with the internet and social networks contributing directly to vaccine hesitancy movements, since anti-vaccine campaigns generally make use of infodemic ecosystems to propagate antivax discourses.

The goal of the current proposal is therefore to map out these antivax discourses on Twitter in Brazil. To do so, we employed Lexical Multidimensional Analysis as a method for identifying the major lexical dimensions in the corpus (Berber Sardinha, 2019; Berber Sardinha and Fitzsimmons-Doolan, in press; Clarke, 2019), which correspond to the prevailing discourses around the anti-vaccine movement. Corpus keywords were computed by comparing the counts of each word against a reference corpus comprising 100,000 tweets – the words whose occurrence (in terms of text counts) was statistically larger in the target corpus than in the reference corpus (Biber & Egbert, 2018) were selected for further analysis. The counts of these words were entered in a factor analysis, which returned seven factors as the optimum solution. The factors were interpreted discoursally by examining hundreds of posts with marked factor scores, yielding seven dimensions, which include the following: equating refusal to vaccinate with a defense of individual freedom; disseminating side-effects as scare tactics against the benefits of inoculation; and discrediting health officials and scientists by pointing out contradictions in their discourse, among others.

**Keywords:** anti-vaccine, antivax, vaccine, vaccine hesitancy, health, coronavirus, corpus linguistics, lexical multidimensional analysis, discourse analysis, twitter, social media, internet

[4735] ***“Be one with your external iTunes library, young Padawan.” Star Wars-derived words and constructions in English corpora and dictionaries.***

Christina Sanchez-Stockhammer (TU Chemnitz).

**Abstract.** This paper explores the occurrence of words and constructions derived from the Star Wars franchise in English corpora and dictionaries and the interrelationship between the two. Star Wars is an international pop-culture phenomenon comprising at its core the nine films (three trilogies) of the Skywalker saga and various spin-offs, videogames, books and merchandise. The franchise’s ongoing popularity increases the likelihood of speech events referring to the films and their content, but in addition, characteristic words and constructions from the Star Wars universe have also been entering the English language and thus making a transition from the screen to the real world in the past years. The recent inclusion of several Star Wars-derived entries in the Oxford English Dictionary (OED) testifies to their increasing establishment.

This paper presents a corpus study (Sanchez-Stockhammer, to appear) that investigates in what contexts “Jedi”, “Padawan”, “lightsabre” (with spelling variants), “Yoda” and the construction “to the dark side” occur in four corpora of present-day English (COCA, COHA, BNC, BNC Spoken 2014). Random samples of up to 100 occurrences per item were assigned to one of five categories based on the type of reference to Star Wars: e.g. direct reference to the films (as in “‘I’d watch it just for the LIGHTSABER battles.”) vs. innovative uses (e.g. “A developer crosses over TO THE DARK SIDE and learns marketing”; both COCA).

The findings suggest that a considerable proportion (over one third of the annotated tokens) display innovative uses, thus reaching a high level of integration into the English language.

In line with the conference’s thematic focus on lexicography, the final part of the paper investigates to what extent these corpus findings are reflected in the corresponding OED dictionary entries’ definitions and quotes and how Star Wars films and texts are used as lexicographical resources (e.g. for quotes).

**References**

Sanchez-Stockhammer, Christina. Accepted (Feb. 2023). The impact of Star Wars on the English language: Star Wars-derived words and constructions in present-day English corpora. To appear in Linguistic Vanguard.

**Keywords:** Star Wars, establishment, corpora, lexicography, Oxford English Dictionary

[4780] ***The language of men's liberation: A corpus-assisted Positive Discourse Analysis.***

Alexandra Krendel (Lancaster University).

**Abstract.** In this talk, I establish the topics which are central to an online community of pro-feminist men on the content aggregation site Reddit known as r/MensLib (short for men’s liberation). r/MensLib state that their purpose is “building a new dialogue on the real issues facing men through positivity, inclusiveness, and solutions-building” (Reddit, n.d). The study of online pro-feminist men’s communities is relatively uncommon compared to online anti-feminist men’s communities such as those associated with the manosphere (e.g. involuntary celibates). Drawing on the concept of Positive Discourse Analysis (Martin, 2004), I argue it is important to establish how men’s issues can be discussed in a progressive and positive manner, so such alternative discourses can be promoted and critiqued.

I collected the 50 most upvoted posts of all time and their associated comment threads from r/MensLib on the 28th November 2022. This corpus totalled 925,153 words. I then used SketchEngine to conduct a keyword analysis using ELEXIS English Web 2020 corpus as my reference corpus, to capture the words which were central to r/MensLib compared to general online language. The 50 most statistically significant keywords were inductively categorised into semantic domains, and concordance lines were used to establish how these keywords were used in context.

The most prevalent keyword categories included: gender roles (masculinity, manly), the LGBTQ+ community (non-binary, bisexual), the manosphere and the alt-right (incel, mra), transphobia (terf, transphobic) and sexism (misogyny, sexist). r/MensLib framed themselves as allies of the LGBTQ+ community, they consciously distinguished themselves from the manosphere and they discussed issues which have been historically associated with feminist movements such as access to abortion, body-shaming and ethical issues surrounding pornography.

**References**

Martin, J. 2004. Positive Discourse Analysis: Power, Solidarity and Change. Revista Canaria de Estudios Ingleses, 49: 179– 200.

Reddit. (n.d.). r/MensLib. Retrieved December 12, 2022, from <https://www.reddit.com/r/MensLib/>

**Keywords:** corpus-assisted discourse studies, masculinity, men's liberation, feminism, positive discourse analysis

[4928] ***Balance and Representativeness in the Compilation of the National Corpus of Irish (CNG).***

Mícheál J. Ó Meachair (Fiontar & Scoil na Gaeilge, DCU), Úna Bhreathnach (Fiontar & Scoil na Gaeilge, DCU), Brian Ó Raghallaigh (Fiontar & Scoil na Gaeilge, DCU) and Gearóid Ó Cleircín (Fiontar & Scoil na Gaeilge, DCU).

**Abstract.** In this paper we report on the feedback loops we have used when engaging in the iterative process of building a balanced and representative national corpus of Irish.

We present the challenges we have encountered, and their solutions, when trying to represent as many voices as possible from a community where a relatively small number of people publish and/or broadcast in the language.

The data collection that has been completed in the first 18 months of this project was done with a view to working towards the domain/genre specifications of the Written BNC 2014 (Brezna, Hawtin, McEnery; 2021). We have, however, added subgenres to this specification that have been deemed important to the Irish language because of their use in education, their importance to the community from a heritage perspective, or their cultural importance. For example: Lúibíní (ditties/songs with turn taking), Sean Nós (singing in the old style), Oral Fiction, Oral History, Plays or Stage Dramas, among others.

Dialectal data has long been of interest to scholars, language learners and other cohorts of the Irish-language community (e.g. Wagner, 1958; O’Rahilly, T. F. 1932), and dialectal labels were included in the metadata of the New Corpus for Ireland (Kilgarriff, et al. 2006). The categorisation of dialectal data is challenging at scale, therefore, we offer a computational solution to this challenge that utilizes machine-learning techniques.

In the first of our feedback loops we conduct wordlist analyses of publication types with a view to monitoring changes, or lack thereof, as the data are changed. These analyses are being used to monitor the impact of balancing changes a posteriori, rather than being done a priori with a view to deliberately affecting word frequencies.

The second of our feedback loops involves metadata analyses whose purpose is to monitor word counts associated with a broad spectrum of metadata; such as author, gender-based designations, subgenre, and so on. When new data have been collected from smaller sub-genres, for example, they displace data from the largest sub-genre until the proportions of a sub-genre are in proportion or until no additional data can be found.

**Keywords:** representativeness, balance, corpus design, sampling, minority language

[4986] ***The language of fashion from a multi-dimensional perspective.***

Katherine Oliva Ortolani (Pontifical Catholic University of Sao Paulo).

**Abstract.** Although fashion influences global society in different ways, this domain received little attention from a corpus perspective. According to Moeran (2004), fashion ‘could not exist without language.’ Most studies looking at the verbal language of fashion focus on topics such as terminology (Koester & Bryant, 1991), cross-language borrowing (Balteiro & Campos, 2012), and social semiotics (Barthes, 1983, 2013). There is a lack of studies on fashion from a discourse perspective, with discourses being ‘ways of looking at the world, of constructing objects and concepts in certain ways, of representing reality’ (Baker & McEnery, 2015, p. 5). To fill this gap, this paper presents a study whose goal is to detect the major discourses in the domain of fashion in

English texts. A register-diversified corpus of fashion-related texts was compiled, including a wide range of varieties, namely newspaper and magazine articles, television programs, social media posts, blogs, documentaries, academic texts, fiction books, and films. The corpus was analyzed using Lexical Multi-Dimensional analysis (LMD; Berber Sardinha, 2019, 2021; Berber Sardinha & Fitzsimmons-Dooley, in prep; Fitzsimmons-Dooley, 2019, forthcoming; Clarke et al. 2022), which derives from the MD analysis framework (Biber, 1988). LMD Analysis enables the identification of the lexical parameters of variation across the texts, which are detected statistically using Factor Analysis. Methodologically, the corpus was tagged for POS and lemmatized, the most frequent lemmas were selected and entered in a factor analysis, and the resulting factors were interpreted as dimensions. Each text in the corpus was scored on each of the dimensions, and the scores were compared across the different registers. The dimensions, which will be introduced and illustrated in the presentation, correspond to the major discourses on which the domain of fashion relies for a range of business and social goals, such as marketing, social influencing, and trendsetting.

**Keywords:** Corpus linguistics, Fashion, Discourse, Lexical Multi-Dimensional Analysis

[5059] ***Analysing and visualising free-text comments: A corpus-based toolkit.***

Dawn Knight (Cardiff University), Paul Rayson (Lancaster University), Nouran Khallaf (Lancaster University), Steve Morris (Cardiff University), Mo El-Haj (Lancaster University) and Ignatius Ezeani (Lancaster University).

**Abstract.** Free-text qualitative comments (e.g. from surveys and questionnaires), pose a particular challenge to a range of companies/institutions, who may not have the expertise to analyse these comments with ease. Following the Welsh Language (Wales) Measure (2011), survey respondents in Wales should be given the opportunity to respond to surveys in English or Welsh, posing even more of a challenge when analysing the resultant data, if adequate Welsh language expertise do not exist. Although a range of sophisticated tools for the analysis of text-based data are already available, many of these tools are not necessarily affordable, quick and easy to use, and/or accessible to non-expert user, nor do they fully support the task of systematically processing free-text responses in Welsh and English.

This presentation reports on the developments of a unique open-source online free-text analysis tool that has been designed to respond to this need: FreeTxt. Funded by the AHRC, and co-designed/co-developed in collaboration with project partners National Museum Wales, National Trust Wales and Cadw, FreeTxt is a unique corpus-based analysis toolkit that is designed to enable the quick and easy analysis of English and Welsh language data, and to engage new user groups with corpus-based methods in new ways. In this presentation we will:

- underline the importance of user feedback, and articulate the key challenges of tackling such data,

- present a novel corpus-based approach to the analysis of FreeTxt data, which can be adapted to multiple languages and contexts,

- outline the key functionalities of the tool, which include: KWIC, POS tagging, semantic tagging, summarisation and sentiment analysis utilities, a novel n-gram frequency tool, text visualisation and multilingual support, and

- provide a demonstration of an early version of the FreeTxt tool in action, using data from survey responses and online feedback forums.

**Keywords:** Free-text, Welsh, Toolkit, Multilingual, Visualisation

[5153] ***Authorial presence across Medicine RA sections: a corpus-based study of self-mentions.***

Heng Gong (University of Auckland) and Michael Barlow (The University of Auckland).

**Abstract.** Based on a specialised corpus consisting of 1,533 IMRD-structured RAs in Medicine, this study investigated both the frequency and functional distribution of personal, impersonal, and mixed self-mention markers across four sections of the articles. We found that personal and mixed self-mentions were generally preferred over the impersonal forms. Among the IMRD sections, the personal self-mentions show a different flow of authorial presence compared with the other two forms, and the former were especially salient in M while the others were more frequently applied in D. Overall, the personal group is most frequent self-mention device in each section. The Correspondence Analysis further showed that there existed evident textual colligations between individual self-mention markers and the IMRD sections. It revealed that the personal self-mention *we* was clearly associated with M; whereas both the mixed self-mention our+noun and the impersonal forms -- this study, the current study, and the present study -- were more favoured in D. An analysis of the discourse functions of three typical self-mention (*we*, our+noun, this study) devices in terms of their verb collocates further revealed that the deployment of such metadiscursive resources across MEDI RA sections was not arbitrary but followed specific patterns. We identified that the plural first-person pronoun *we* was the prevailing self-mention device in MEDI RAs, showing a more significant variation of both verb collocates and discourse functions than the others. Another Correspondence Analysis revealed that each type of self-mention had its preferred discourse functions and associated authorial identities.

**Keywords:** authorial presence, Medicine RA sections, self-mentions, verb collocates

[5213] ***The hierarchy of web pages: accounting for contents’ accessibility in keywords analysis.***

Matteo Di Cristofaro (Università di Pisa).

**Abstract.** The use of keywords for Corpus-Assisted Discourse Studies approaches is an established practice to investigate similarities and differences in the use of words across different collections of real occurrences of language. At its core, keywords analysis rests upon the use of statistical measures that take into account the frequency of words contained in two or more (sub-)corpora, with more recent studies suggesting the inclusion of the dispersion of words as a further element to account for keywords’ keyness. The application of keywords analysis on contents extracted from webpages however poses a further challenge, namely that the hierarchy with which contents are organised directly impacts the fruition of the contents themselves (cf. Djonovan 2007). In order to account for this impact, we propose the integration of content and organisation features to the study of keywords, providing an additional layer of information able to capture the pragmatic and interactional value the web page hierarchy has when browsing web contents. Utilising data collected as part of the project “Communicating transparency”, the method proposed employs practices from network analysis (e.g. ‘shortest-path’ and centrality) as a mean to account for the dispersion of language inside the hierarchical network set out by the contents creators. The results complement traditional keywords analysis with a metric able to capture potential exposure to words that user experience when browsing the website contents, along with the importance the creators give to the language used.

**References**

Djonov, Emilia. ‘Website Hierarchy and the Interaction between Content Organization, Webpage and Navigation Design: A Systemic Functional Hypermedia Discourse Analysis Perspective’. Information Design Journal 15, no. 2 (10 August 2007): 144–62. <https://doi.org/10.1075/idj.15.2.07djo>.

**Keywords:** keywords, corpus linguistics, network analysis, web for corpus

[5255] ***Gender on Dimensions of Linguistic Variation. Making Sense of Non-significance.***

Marianna Gracheva (Northern Arizona University).

**Abstract.** The study explores the relationship between social and situationally determined linguistic variation in a corpus of business memos, image descriptions, emails, academic essays, evaluations, interviews, and text messages produced by 19-23-year-old university students in the UK (corpus of ‘100 Idiolects’, Heini, Kredens & Pezik, 2021). Specifically, the study examines gender variation with regard to general communicative patterns in these registers. Thus, the study first identifies the underlying patterns of linguistic variation among texts through a multidimensional analysis and produces linguistic dimensions of ‘Online Oral Elaboration vs. Information Density’, ‘Abstract vs. Concrete Discourse’, ‘Others-oriented vs. Self-oriented/ Interactive Discourse’, and ‘Evidence-based Stance’. The study then investigates whether gender predicts linguistic variation on these dimensions within each register.

The results show that gender groups do not differ in ways they navigate the dimensions in any of the registers. Rather, both gender groups employ a range of their linguistic resources in each register. The study offers possible explanations for this outcome and further tests whether gender differences lie elsewhere. The study observes several noteworthy patterns. First, in contrast to specific functions of individual linguistic features, previously found to reflect gender differences in language, general communicative patterns based on functional feature co-occurrence revealed by multidimensional analysis are navigated the same way by men and women. Second, when other linguistic features are examined, register appears to play a part in gender differences. These observations are in line with and are discussed with reference to some previous gender research, which highlights the role of situation as well as other social factors in the observed results.

**Keywords:** register, gender, multidimensional analysis, communicative patterns in discourse

[5333] ***On very few people and très peu de gens – translating quantity modification with English (a) few and French (un) peu.***

Lobke Ghesquière (University of Mons (UMONS)) and Jesse Marion (University of Mons (UMONS)).

**Abstract.** In contrast to the extensive body of research into quantifiers (e.g. Barwise & Cooper 1981; Langacker 1991, 2008, 2016, 2017; Benninger 1999; Radden & Dirven 2007), the modification of those quantifiers (e.g. almost all dogs or very few people), i.e. quantity modification, has started receiving attention only recently (e.g. Njende et al. 2015, 2017; Davidse 2018; Author 2021, 2022).

To contribute to the understanding of quantity modification from a cognitive-linguistic perspective, this corpus-based synchronic study looks into the translation of two absolute quantifiers undergoing quantity modification, English *(a) few* and its French counterpart *(un) peu*. The study will verify previous findings which said that absolute quantifiers not only pair with scalar (up- and downscaling) modifiers, but also with compromisers, emphasisers and focusers (Njende et al. 2017).

The translation data used in this study come from the Europarl Direct corpus, which totals approximately 50 million words per language collected from 1996 to 2010 (Koehn 2005; Cartoni & Meyer 2012). The written comparable 29-million-word Yahoo-based Contrastive Corpus of Questions and Answers was used for the monolingual study (De Smet 2009). Its collection period ranges from 2006 to 2009. Particular attention will go to the types of modifiers the items under study take and the impact these combinations have on the translation outcomes.

Preliminary research has already shown that quantity modification is often lost in English to French translation, with translators opting for syntactically non-equivalent constructions, e.g. *very few people* becoming *un petit nombre de gens*.

This study will therefore allow a better understanding of the modification of absolute quantifiers in monolingual data as well as in translation, and will inventory the choices translators make when confronted with modified *(a) few* and *(un) peu*.

**Keywords:** corpus linguistics, cognitive linguistics, quantifiers, quantity modification, (a) few, (un) peu, translation study, monolingual study, Europarl Direct, Yahoo-based Contrastive Corpus of Questions and Answers, English, French

[5366] ***Can a corpus analysis of “such… as” constructions in 18th century American English facilitate interpretation of the U.S. Constitution’s Appointments Clause?.***

Haoshan Ren (Georgia State University), Clark D. Cunningham (Georgia State University) and Ute Römer (Georgia State University).

**Abstract.** The Appointments Clause of the U.S. Constitution requires Presidential nomination and Senate confirmation for “all Officers of the United States” unless a provision called “the excepting clause” is applicable, which reads: “the Congress may by law vest the appointment of such inferior officers, as they think proper, in the President alone, in the courts of law, or in the heads of departments.” The U.S. Supreme Court interprets “inferior officers” as meaning “subordinate to officials who have been confirmed by the Senate,” even though it is not clear that the text supports such an interpretation. The Supreme Court uses this interpretation to deny Congress the power to create positions in the executive branch that have independence from the authority of Senate confirmed officials.

Our linguistic analysis raises doubt about whether the Supreme Court’s interpretation is consistent with the meaning of the Constitutional text, which embeds “inferior officers” within the well-studied construction “such + noun phrase + as.” We used data from the Madison Papers, a 10.8-million word sub-corpus of the Founders Online database (founders.archives.gov), consisting of James Madison’s public and private correspondence, public actions and speeches, and political writings from 1751 to 1836, to investigate how the “such + adj + noun + as” construction was used in Founding Era American English.

Concordance lines containing the target construction were extracted and then manually sorted into three categories, depending on the function of each example’s as-phrase: (1) as-phrase as a descriptive qualifier, (2) as-phrase as an exemplar, and (3) as-phrase as a discretionary qualifier. We conclude that the excepting clause should be interpreted as using a discretionary as-phrase, which suggests that the Constitutional text empowers Congress, not the courts, to make the final decision as to whether to authorize appointment of officials who have independence from Senate-confirmed officers.

**Keywords:** Law and corpus linguistics, Legal discourse analysis, Constructions, Historical corpus analysis, Corpora and context, Collocation and meaning in context

[5383] ***Business discourse in the Irish technology sector: a corpus-based investigation of spoken language in International Virtual Teams (IVTs).***

Gail Flanagan (University of Limerick).

**Abstract.** This study introduces the International Virtual Team Corpus (hereafter the IVT corpus). This corpus currently consists of 80,000 words of online business meetings with plans for expansion in 2023 to 150,000 words. The reason for selecting the technology sector is due to its significance as an employer in Ireland with 16 of the top 20 global technology firms and 9 of the top 10 US Information and Communications Technology (ICT) companies locating strategic operations in Ireland (Technology Skills 2022 report). The present researcher’s professional experience also confirms that for this sector, virtual meetings were the norm, even pre-pandemic. This is further supported by the results of a survey of technology sector professionals (N=113) in the initial research phase where it was determined that Irish participants spend an average of 3.5 hours a day in virtual meetings, conversing with up to 13 different nationalities in a single meeting. In turn, these findings align with the overarching goal of the research study which is to develop a set of International Virtual Communication (IVC) Principles, scaffolding the creation of experience-based intercultural communication training for higher education and professional learners.

This paper presents both quantitative and qualitative corpus-based results. Significant keyword clusters and patterns have been identified which provide a window into the pragmatic dimensions of business-oriented IVT meetings in this sector. Focus has been given to the spoken language nuances of online meetings, when compared to face-to-face meetings, through a comparison with the Cambridge and Nottingham Business English Corpus (CANBEC). Concordance lines and collocates from both corpora have been analysed to support the quantitative findings. Furthermore, Conversation Analysis (CA) techniques have been used to deep dive into selected corpus extracts, mining the rich pragmatic layers of phatic communication in established IVTs. Some implications and applications for work-based virtual communication are explored in the closing discussion.

**Keywords:** Engineering Technology Discourse, Intercultural Communication, Online Communication, Spoken Business Discourse, Virtual Team Meetings

[5446] ***Examining corpus prototypicality and keyness beyond the lexical level: Experiments with ProtAnt.***

Nicholas Smith (University of Leicester), Laurence Anthony (Waseda University), Sebastian Hoffmann (University of Trier) and Paul Rayson (Lancaster University).

**Abstract.** For linguists working with corpora, a common difficulty after quantitative analysis is deciding which texts to select for follow-up, close analysis, without arousing suspicion of ‘cherry-picking’. The ProtAnt tool (Anthony & Baker 2015) provides a major boost in this respect. Building on a now well-established tradition of corpus keywords analysis (since Scott 1997), and the association between prototypes and frequency of instantiation (e.g., Rosch 1975, Gries 2003), ProtAnt ranks the texts in a target corpus from most to least prototypical according to the number of keywords they contain. ProtAnt’s capabilities have been increasingly exploited in text/discourse analysis (e.g., Levon 2016, Bednarek and Caple 2017, Price 2022), but to date, all such studies have been confined to traditional lexical-based keywords, rather than keywords generated at other linguistic levels such as parts of speech (POS), semantic domains, and speech acts. The current paper seeks to address this research gap, posing the question: How successfully can ProtAnt identify prototypical and outlier texts in corpora at various non-lexical linguistic levels?

We address this question through a series of experiments. Results show that ProtAnt is able to use key POS-tags to identify stylistically prototypical texts in registers of the American AmE06 corpus and also flag outlier texts that have been artificially included from another register. Using semantic tags, outliers are identified with still higher success. Other results show that speech act tags (in SPICE-Ireland) yield more mixed results.

On the whole, non-lexical key items are able to complement those at the lexical level in profiling texts, with success seemingly affected by the granularity of the tags, accuracy of the linguistic annotations, and degree of specialization of the register. We discuss the theoretical and practical implications of our work in areas such as grammar, stylistics, discourse analysis, and data-driven learning (DDL).

**References**

Anthony, L. & Baker, P. (2015). ProtAnt: A tool for analysing the prototypicality of texts International Journal of Corpus Linguistics 20(3): 273-292.

Bednarek, M. & Caple, H. (2017). The Discourse of News Values: How News Organizations Create Newsworthiness. Oxford: Oxford University Press.

Gries, S. (2003). Towards a corpus-based identification of prototypical instances of constructions. Annual Review of Cognitive Linguistics, 1:1-27. DOI: 10.1075/arcl.1.02gri

Levon, E. (2016). Qualitative analysis of stance. In Baker. P. & Egbert, J. (eds.) Triangulating Methodological Approaches in Corpus Linguistic Research. London/New York: Routledge.

Price, H. (2022) The Language of Mental Illness: Corpus Linguistics and the Construction of mental illness in the Press. Cambridge: Cambridge University Press.

Rosch, E. (1975). Cognitive representations of semantic categories. Journal of Experimental

Psychology: General, 104(3): 192–233. DOI: 10.1037/0096-3445.104.3.192

Scott, M. (1997). PC analysis of key words - and key key words. System 25(2): 233-245.

**Keywords:** text prototypicality, keywords, parts of speech, semantic domains, speech acts

[5469] ***Evaluating collocation in spoken dialogic corpora.***

Robbie Love (Aston University), Isobelle Clarke (Lancaster University) and Mark McGlashan (Birmingham City University).

**Abstract.** The study of collocation is a fundamental approach in the corpus linguistics toolkit. Many studies use the ‘collocation window’ method, which measures collocation within a range, for example, five words to the left (5L) and right (5R) of the given node word (Gablasova et al., 2017: 158). This approach is facilitated by mainstream corpus tools – for example Sketch Engine – that allow the user to define the collocational span according to their research interests. At a span of 5L and 5R, the tool searches for collocational patterns within strings of up to 11 tokens in length (five either side of the node, plus the node). For corpora where each text is individually authored, collocation windows occur within texts. Yet, when files within corpora consist of multiple texts authored by multiple authors, some collocation windows occur across text boundaries.

Dialogic spoken corpora are a case in point, tending to comprise transcripts of conversations between two or more speakers. In this context, each corpus text contains many utterance boundaries as the speakers produce turns. In casual conversation, speaker turns may routinely amount to just a few tokens in length; therefore, the collocation window method – when applied to dialogic corpora – routinely searches windows that span across utterance boundaries. The outcome of this is that collocations ‘co-produced’ by two speakers (producing one word each of a collocation pair) are in no way distinguished from collocations produced solely by individual speakers. Consequently, in a case study of collocation in the Spoken BNC2014, this talk shows the effect of restricting the collocational span to utterance boundaries in comparison to spans of 5L and 5R.

**Reference**

Gablasova, D., Brezina, V., & McEnery, A. M. (2017). Collocations in corpus-based language learning research: identifying, comparing and interpreting the evidence. Language Learning, 67(Suppl. 1), 155-179. <https://doi.org/10.1111/lang.12225>

**Keywords:** collocation, spoken corpora, conversational corpora, dialogic corpora, collocation windows

[5584] ***The emergence of 'publicly': using a historical corpus to investigate an orthographic anomaly.***

John Williams (University of Portsmouth).

**Abstract.** The adverb 'publicly', in its standard spelling, breaks the rule whereby adjectives ending in –ic form their corresponding adverb by the addition of –ally rather than simply –ly (cf. 'basic' --> 'basically'). There are no other common, standard -icly adverbs.

Our focus is on the historical emergence of 'publicly (via 'publickly'). This was investigated principally by means of the English Historical Book Collection (EHBC) corpus (<https://www.sketchengine.eu/historical-collection-eebo-ecco-evans/>), with just under one billion tokens covering the period 1470-1829.

A regexp search designed to capture all potential spelling variants of ‘publicly’ produced 25,204 concordance hits (52 word forms). Infrequently attested till C16, the word gains ground from the 1550s onwards, peaking in C17 at 27.9 tokens per million (higher than in most modern corpora).

Examples from several sources suggest that the rise of ‘publicly’ in C16 and C17 was connected with the distinction between public and private worship that informed religious disputes during those centuries (Jensen, 2012, <http://beingbess.blogspot.co.uk/2012/03/religious-policy-under-elizabeth-i.html>; Haldén, 2013).

Following Oakes (2020), a chi-square test was used to confirm a strong collocational association between ‘publicly’ and ‘privately’ (any spellings) in EHBC (window ± 10; X2 (1, N = 25204) = 26857417, p < .05). ‘Publicly’ also manifests a strong association with an arbitrarily chosen set of words from the semantic field ’religion’ (window ± 10; X2 (1, N = 25204) = 180053, p < .05.). To counter the objection that ‘everything was about religion in those days’, 6.7% of instances of ‘publicly’ were found to ‘attract’ at least one of our ‘religion words’ in their context, as opposed to 0.06% for all other node words together.

Our contention is that the spelling ‘public[k]ly’ became established as a result of orthographic regularization with ‘privately’ (cf. ‘bizzie lizzie’, ‘rack and ruin’). The adjective ‘private’ simply adds -ly, and ‘public’ follows suit. In a range of modern corpora, despite some weakening of the association, the spelling ‘publicly’ consistently attracts ‘privately’ more frequently than does the non-standard variant ‘publically’, suggesting that the motivation for the -icly spelling is still present in vestigial form.

**Other References**

Haldén, P. (2013). Fundamental but not eternal: The public-private distinction [...] in Western political thought. Small Wars And Insurgencies, 24(2), 211-223

Oakes, M. (2020). Statistical significance for measures of collocation strength (WP3). In Pastor, G.C. & Colson (Eds.). Computational Phraseology. John Benjamins, pp.189-206

**Keywords:** lexicology, orthography, historical corpora

[5710] ***Causes, effects, and human-nature relations: A transitivity and inter-institutional analysis of UK climate change discourses.***

John Currie (Østfold University College and University of Gothenburg), Ben Clarke (University of Gothenburg) and Daniel Lees-Fryer (Østfold University College).

**Abstract.** It has been argued that a central cause of climate change and the environmental crisis is the dichotomisation of humans and nature, which partly characterises the Anthropocene (e.g. Merchant, 2020). Viewing ourselves as separate from nature allows us to participate in environmentally destructive behaviour, waging ‘war’ on nature. The irony is that humans are effectively waging war on themselves. We have shown that this war metaphor is increasing in use in UK parliamentary discourse, as political decision-makers create a narrative of People vs Climate Change, while seemingly failing to recognise human behaviour as a cause of climate change (Authors, 2022). Just as environmental damage can more easily occur if we view ourselves as separate from nature, so too will it be difficult for humans to ‘fight’ against anthropogenic climate change while acknowledging human behaviour as a cause.

Here, we present the results of a 2015-2022 UK inter-institutional analysis (Anderson, 2014), which examined how causes and effects of climate change are represented in parliamentary, scientific, activist and news media corpora. First institution-specific definitions of causes and effects were retrieved; second, returns were categorised according to our own thematic typology (e.g. causes: Human effect on land e.g. ‘deforestation’, effects: Biodiversity e.g. ‘mass extinction’), where already there were institutional differences; third, a number of transitivity analysis patterns (Halliday & Matthiessen, 2004) were identified at a corpus level, based on verbs frequently collocating with these causes and effects. For example, metaphorised occurrences of material transitivity verbs occurred more frequently in the activist corpus (e.g. ‘halt biodiversity loss’) where in the scientific corpus their literal counterparts were typically found (e.g. ‘decrease species richness’). We interpret these findings in terms of climate mitigation and adaptation (e.g. Moser 2014) and the extent to which focus is on ‘tackling’ human-specific vs nature-specific effects and human-specific vs generalised causes.

**References**

Anderson, A. (2014). Media, environment and the network society (Palgrave studies in media and environmental communication). Basingstoke: Palgrave Macmillan.

Authors. (2022). Fighting talk: The use of the conceptual metaphor climate change is conflict in the UK Houses of Parliament 2015-2019. Journal of Language and Politics, 21(4), 589-612.

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2004). Introduction to functional grammar. 3rd ed. London: Arnold.

Merchant, C. (2020). The Anthropocene and the humanities : From climate change to a new age of sustainability. New Haven: Yale University Press.

Moser, S. C. (2014). Communicating Adaptation to Climate Change: The Art and Science of Public Engagement When Climate Change Comes Home. WIREs Clim Change, 5, 337–358.

**Keywords:** climate change, climate adaptation and mitigation, human-nature relations, transitivity analysis, metaphor

[5775] ***Topic evolution of a political genre: analysis and visualization of a corpus of Presidential Messages to the Congress in Chile (1832 - 2022).***

Riva Quiroga (Pontificia Universidad Católica de Chile).

**Abstract.** Since 1832, the annual presidential message to the Congress in Chile has registered the political priorities of each of the presidents, the challenges faced by the country, and the key topics that have been considered relevant to discuss. Despite its social and political importance to the country, this genre has not been studied from a diachronic point of view. The only studies that exist have focused on lexico-grammatical or discursive aspects of specific presidents (e.g.; Soto & Zenteno, 2010, among others) or the way in which certain topics have been covered during specific political periods (e.g. Martínez, 2019, among others).

In this context, this talk will present the results of a research project that aims to advance in the description of this genre and to understand the changes that have taken place throughout its history. Specifically, we will present results regarding the diachronic evolution of the topics covered in each speech. To this purpose, each of the 188 speeches was labelled according to the topics covered in each of its sections (e.g., External Affairs, Justice, Education, Healthcare, First Nations, etc.) and analyzed and visualized using R (R Core Team, 2020) in two key aspects: 1) the frequency and distribution of each of the topics from 1832 to 2020; 2) the patterns of vocabulary change and stability over time inside each topic. Results show not only that political priorities have changed throughout the years but also that the way words are used (or not used) to talk about some topics has evolved. The talk might be of interest to people working with similar political genres, diachronic corpora, or those interested in data visualization.

**References**

Martínez, N. (2019, January 9-11). “‘Igualdad’ y ‘Desigualdad’ en los mensajes presidenciales de la postdictadura (1990 -2013): discurso político e imaginarios sociales [‘Equality’ and ‘Inequality’ in the presidential messages of the Post-Dictatorship period (1990-2013): political discourse and social imaginary]. [Conference presentation] XI Conference of the Latin American Association of Discourse Studies (ALED), Concepción, Chile.

Soto, G. & Zenteno, C. (2010). La relación en el discurso entre el pretérito perfecto compuesto y las expresiones evaluativas: Análisis de una cuenta pública presidencial [The relationship between present perfect and evaluative expressions: analysis of a presidential message] . Signos, 43(73), 333-354. http://dx.doi.org/10.4067/S0718-09342010000200007

R Core Team (2022). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Austria. https://www.R-project.org/.

**Keywords:** Diachronic specialized corpora, Political Genres, Spanish-language corpora, Data Visualization

[5901] ***Towards a systematic comparison of keyword categorisations.***

Nathan Dykes (Friedrich-Alexander-Universität Erlangen-Nürnberg) and Stephanie Evert (Friedrich-Alexander-Universität Erlangen-Nürnberg).

**Abstract.** Manual annotation plays a central role in corpus linguistics, computational linguistics, social sciences, and many other disciplines. This also applies to Corpus-Assisted Discourse Studies (CADS), where researchers typically start from a list of keywords or collocations that are grouped into ad-hoc categories in order to identify discursive patterns.

The details of this grouping process have largely been neglected in methodological discussions of CADS – likely because evaluation studies usually work on the assumption that the goal is either to reach a consensus and develop annotation guidelines, or to evaluate annotation quality according to an existing gold standard by calculating inter-annotator agreement. However, the very notion of a “universal truth” in the form of a gold standard is hardly compatible with discourse studies, where the perspective of the researcher is taken to be an important part of the interpretation. Understanding how different researchers annotate the same data and reach their categorisations would reveal the similarities and differences of their perspectives and contribute to developing the methodological foundations of CADS.

This paper explores a case of researcher triangulation in CADS, where two researchers were given the same lists of 200 keywords each for two German Telegram channels operated by conspiracy theorists (in comparison to a general reference corpus). The annotators worked independently and came up with separate category schemes, which we aim to compare. The goal of our study is to improve the understanding of what different perspectives the complementary annotations yield on the data, and how they relate to one another. Similarities and divergences are investigated through exploratory visualisation. Despite the researchers’ similar background, the analysis suggests that differences in granularity of categories, in the degree of interpretation attached to the labels and in their abstractness could significantly influence the perspective provided on the data.

**Keywords:** keywords, annotation, corpus-based discourse analysis, corpus methodology

[5980] ***Uncovering indigeneity in a corpus of colonial Quechua.***

Chad Howe (University of Georgia).

**Abstract.** In situations of social and political inequality, particularly stark in colonial situations like that of Latin America, the impacts of long-term subjugation are codified in the texts of minority languages. This project assumes the perspective that textual analysis can be realized through the careful observation of specific patterns of language use and seeks to bring digital humanities-informed methods to the analysis of colonial texts produced in Quechua, an Andean language with nearly 10 million speakers in South America (Coronel-Molina 2015). To test this approach, we are working with three colonial dramas (16th – 19th centuries) written in Quechua: Ollantay, El probre más rico, and Usca Paucar.

Translations of Spanish works were performed in Quechua with Peruvian playwrights later producing ostensibly ‘original’ works in Quechua, the objectives of which were, according to Trucharte (2014:59), complex and could be viewed as representing a contradiction between, on the one hand, the continued expansion of Christian beliefs among indigenous communities and, on the other, the rise in influence of a creole urban-elite who could read and write in Quechua (Itier 1991).

Further complicating this issue is the fact that authors of colonial Quechua dramas, and specifically those mentioned above, were very often anonymous, and the origin of the stories themselves were unknown.

The primary research questions underlying this project concern the authorship of these texts and the extent to which these works represent legitimately indigenous products, both in terms of the authorship and the narrative content (Brokaw 2006). Our approach seeks to address the issue of authorship in these texts, using a novel approach through the application of Natural Language Processing (NLP) tools to leverage language-specific patterns and comparisons to materials of known linguistic and cultural provenance. This work is based on the first fully annotated corpus of colonial Quechua (Howe, Hale, and Peterson 2022) and leverages NLP tools (e.g., Rios 2015) to discover grammatical patterns that, we argue, serve as indices of indigeneity in these documents (e.g. word order). More generally, this research contributes to the development of computational tools in the analysis of under-resourced languages.

**References**

Brokaw, G. 2006. Ollantay, the Khipu, and Eighteenth-Century Neo-Inca Politics. Bulletin of the Comediantes. 68:31-56.

Coronel-Molina, S. M. 2015. Language Ideology, Policy and Planning in Peru. Berlin: De Gruyter.

Howe, Chad, John Hale, and Brent Peterson. 2022. Corpus of colonial Quechua. TEI-compliant xml. Quechua text available in the TAPAS Project.

Itier, C. (1991). Lengua general y comunicación escrita: cinco cartas en quechua de Cotahuasi. Revista Andina. 17:65-107.

Rios, A. 2015. A Basic Language Technology Toolkit for Quechua. Doctoral Dissertation. University of Zurich.

Trucharte, M. 2014. El uso de las lenguas indígenas en el teatro colonial. El ejemplo del quechua. Kipus: Revista Andina de Letras. 35:49-61.

**Keywords:** colonial texts, indigeneity, Quechua, authorship

[6055] ***Teaching spoken features with the American Movie Corpus: a task-based approach.***

Valentina Morgana (Catholic University, Milan).

**Abstract.** In the last ten years, corpus linguistic researchers have pointed out the need to teach spoken grammar features explicitly as this allows learners to speak authentic English and therefore to sound more natural (Biber & Reppen, 2002). However, the teaching of conversational skills in the EFL secondary school classroom has been mainly focused on meaning-making and fluency activities. This study aimed to raise students’ awareness of spoken features using the American Movie Corpus (AMC) (www.americanmoviecorpus.net, see also Forchini, 2021), by adopting a task-based approach in a secondary school context.

Forty students of English as a foreign language participated in an experimental study which was designed to provide students with strategies to use a language corpus as a resource of spoken grammar features. Following the task-based language teaching cycle, learners were first exposed to the spoken target features in the pre-task by watching short movie clips, as a main task they investigated the use of discourse markers in the corpus using the LancsBox Corpus analysis tool, and in the last stage performed a speaking task similar to the pre-task.

The study followed a mixed method approach using both qualitative and quantitative data. A statistical analysis of pre-, post- and delayed post-test data was conducted to measure learners’ knowledge of discourse markers. Additionally, we measured the effects of data-driven learning (DDL) on speaking task performance in terms of the number of attempts and accurate use of targeted spoken features. Qualitative data include semi-structured interviews with students on learning processes and strategies. Results showed how the use of a DDL approach with movie language facilitated the noticing, learning and awareness of specific spoken features. The AMC itself proved to be an authentic source of real-life examples of spoken English grammar, and could enable learners to acquire more conversational skills earlier in their studies.

**Keywords:** Spoken English, Data-driven learning, Corpora in language teaching

[6088] ***An AI-Assisted Annotation Approach for Non-coders: Beyond Modelling Metadiscourse.***

Wenwen Guan (University of Amsterdam).

**Abstract.** Despite the extensive application of AI to corpus annotation, metadiscourse categorisation still relies heavily on manual endeavor due to its variable forms and multifunctionality. The two traditional procedures, known as the “thin” and “thick” approaches, are either inefficient or do not generate context-sensitive labels (Ädel & Mauranen, 2010).

Faced with the dilemma, this study investigated an AI-assisted annotation approach in an attempt to improve both efficiency and precision. The method is implemented using Prodi.gy, an all-in-one annotation tool with embedded machine learning functions. It is an advantage over conventional corpus tools and NLP methods that all phases can be executed on a single platform without the need for prior coding experience. This study illustrates the workflow of our metadiscourse annotation project, commencing with data preparation and ending with result analysis.

What makes Prodi.gy truly remarkable is its suggester function. In other words, users can train a model with a little amount of manually labelled data, and Prodi.gy will suggest labels for the remaining data. To determine how much the function contributes to the annotation process, we trained models with varying amount of manually labelled data and observed the accuracy of model-suggested labels. Moreover, we evaluated the models’ performance on various categories of metadiscourse, allowing us to zoom into the categories that are difficult to predict. This enabled us to provide specific feedback for annotator training and particular tricky categories for qualitative analysis.

AI-assisted annotation proves to be an excellent compromise between fully manual and automatic annotation. The suggester function greatly reduced manual efforts and produced acceptable accuracy. The approach also extended to annotator training and qualitative analysis. Our workflow is replicable for studies on other functional linguistic subjects like metadiscourse and has the potential to be applied to a variety of linguistic data, such as audio and sign language data.

**Keywords:** metadiscourse, context-sensitive data, AI-assisted annotation, annotator training

[6106] ***Science at Cultural Crossroads: Hedging in Arabic Discourse on Evolution.***

Mohammad Aboomar (Dublin City University).

**Abstract.** This paper investigates hedging in two corpora of Arabic texts on the culturally sensitive domain of evolutionary biology. The paper asks to what extent hedging is preserved in discourse on evolution in Arabic science publications translated and written by professionals, and whether translated and non-translated discourse on the topic display different levels of hedging. The analysis covers materials published between 2016 and 2020 through a parallel corpus of about 200 English-Arabic text pairs aligned on sentence level and an Arabic monolingual corpus of about 300 translated and non-translated texts.

The analysis takes into account three dimensions of the collected texts: genre (features, news, scientific abstracts, etc.); source (Nature, Scientific American, etc.) and mode of publication (print, online). The research builds on Doghaiem’s (2019) investigation of the ideological implications of hedging in culturally sensitive science communication from 3 disciplines (evolution, genetics and sexology) translated from English into Arabic by volunteer popularizers and disseminated on online platforms. Preliminary results suggest a higher adherence to hedging in translated science publications than in online platforms, which warrants an interpretation of the results that goes beyond ideological considerations. The research contributes to the nascent albeit promising area of Arabic corpus linguistics (McEnery, Hardie and Younis, 2019), particularly in Arabic translation studies (Aboomar, 2022).

**References**

Aboomar, M. (2022) ‘لسانيات المتون في الدراسات الترجمية العربية [Corpus linguistics in Arabic translation studies]’, Maalim, 13(2), pp. 69–82.

Doghaiem, A.A. (2019) ‘Facts and falsehoods in activist translations of popular science by Arabs in post-truth era’, mTm, 11, pp. 92–116.

McEnery, T., Hardie, A. and Younis, N. (2019) ‘Introducing Arabic corpus linguistics’, in T. McEnery, A. Hardie, and N. Younis (eds) Arabic Corpus Linguistics. Edinburgh: Edinburgh University Press, pp. 1–16.

**Keywords:** Arabic corpus linguistics, hedging, parallel corpus, science popularization, biological evolution

[6161] ***Movie transcriptions and subtitles: “Two of a kind” for spoken lexicogrammar.***

Marcia Veirano Pinto (Sao Paulo Federal University) and Pierfranca Forchini (Università Cattolica del Sacro Cuore).

**Abstract.** The verbal language of movies can be represented in writing by three text types: scripts, transcripts, and subtitles. The difference in their lexicogrammar derives from their differing purposes, as scripts represent the audio-visual intent of screenwriters and directors, transcripts represent verbatim what viewers hear, and subtitles help viewers follow what is being said.

As linguists our focus is not on movie production, but on the language to which viewers are exposed when watching movies. The aim of this paper is, therefore, to understand how exactly the lexicogrammar of transcripts are similar to/differ from that of subtitles. Given that spoken lexicogrammar is rarely included in educational situations (Hunston 2002, Mauranen 2004, Reppen 2010), despite the importance it has in communication (Bloomfield 1933, Halliday 1987, Biber et al. 1999, Carter and McCarthy 2006) and that movie conversation mimics the linguistic features of face-to-face conversation (Forchini 2012, 2017, 2020), a secondary aim is to verify whether subtitles can also replace spoken language in the classroom when spoken data are not available. To this end, we investigated two corpora of a set of 16 movies. One corpus contained the transcriptions from the AMC (www.americanmoviecorpus.net) and the other the subtitles of those same movies taken from the NAMC (Veirano Pinto, 2014). We submitted the two corpora to the Word and N-gram tools in AntConc 4.1.4 and performed a comparative semi-automatic analysis of both corpora with the help of filters in excel worksheets as well as a script written in shell. As expected, results show that transcriptions are more verbose, but spoken features, such as interjections, discourse markers, fillers, and contractions are also well represented in subtitles. As such, both transcriptions and subtitles are valid sources for research and teaching materials and should be selected according to specific needs and purposes.

**Keywords:** American movies, transcriptions, subtitles, lexicogrammar, conversation features

[6210] ***Linguistic Variation within Registers – Communicative Freedom Meets Situational Constraints.***

Marianna Gracheva (Northern Arizona University).

**Abstract.** The study investigates communicative freedom afforded to language users within registers and shows that the demands of a single situation may be met in a number of different ways.

The study makes use of the corpus of ‘100 Idiolects’ (Heini, Kredens & Pezik, 2021) and examines texts produced by 112 speakers in the registers of business memos, image descriptions, emails, academic essays, evaluations, and interviews. The study first conducts a multidimensional analysis and identifies four general patterns of linguistic variation in the corpus: ‘Online Oral Elaboration vs. Information Density’, ‘Abstract vs. Concrete Discourse’, ‘Others-oriented vs. Self-oriented/ Interactive Discourse’, and ‘Evidence-based Stance’.

On the identified dimensions, the study then examines texts that are representative of each register’s central trend and texts that are on the periphery of the register, i.e., deviate from the central tendency. First, the study observes varying amounts of linguistic freedom permitted by different registers: emails and business memos, for example, allow extensive variation within their scope, while oral interviews on a single topic and evaluations of university resources tend to be quite restrictive situationally and, consequently, linguistically. Importantly, the linguistic variation observed in some registers always corresponds to specific situational parameters that vary across texts. Thus, the study provides empirical evidence for functional correspondence between situation and language within registers as well as between them and illustrates the reciprocal relationship between the situation of use and the communicative choices made by language users within the situation, in which, on the one hand, speakers are constrained by the situation of use and, on the other, their communicative needs may broaden the situational scope of a register.

**Keywords:** register, communicative freedom, intra-register variation, central texts, peripheral texts, multidimensional analysis

[6225] ***Data Driven Learning of Sino-Japanese and Sino-Korean vocabulary.***

Sara Librenjak (York St John University).

**Abstract.** Learners of Japanese and Korean as a second or foreign language have a several common struggles despite studying different and genetically unrelated languages. Vocabulary learning is a challenging task for learners of both Japanese and Korean since both languages have no confirmed cognates, and at times due to a lack of language input.

However, both Japanese and Korean have a large percentage of lexemes based on Chinese language - Sino-Japanese and Sino-Korean vocabulary. Japanese still uses hanzi-derived script kanji, while Korean mostly replaces it with native script hangul, but it is approximated that more than 60% of respective lexicon is derived from Chinese. These lexemes are also more systematic since they're based on logographic script, which can be systematically analysed and sorted by frequency. That makes it relevant for language learners, especially if vocabulary is learned in order by relevance.

In this presentation I propose a method of organising the vocabulary based on frequency of characters that comprise the lexemes, and using that order in language classes. I will look at the frequency in different corpora and compare the differences of Sino-Japanese and Sino-Korean vocabulary depending on the field and register. Lastly, I will present the initial findings of how students react to incorporation of data-driven learning in the language classroom, and share the takeaways that can be used for a larger scale implementation.

**Keywords:** language learning, Japanese, Korean, data-driven learning, learner corpora

[6316] ***US voter rights in translation: Semantic shifts in Spanish renderings of “felony”.***

Matt Riemland (Dublin City University).

**Abstract.** Felons’ voting rights have featured prominently in debates over voter suppression in the United States, particularly in Florida, where a 2018 state constitutional amendment reinstated voting rights to the state’s 1.4 million former felons (Robles 2018). Florida also has a high concentration of Spanish-speaking voters with Limited English Proficiency (LEP), and Latinos are becoming an increasingly significant portion of the Florida electorate (Bergad 2018; Morín et al. 2021). Hopkins (2011) and Parkin and Zlotnik (2011) demonstrate that Spanish-language materials influence election results in monolingual communities. Inadequate translations of voter information may therefore misrepresent voter eligibility for LEP Spanish-speaking former felons and influence election results in Florida.

This study constructs a parallel corpus to investigate how semantic shifts (shifts in meaning) occur in Spanish translations of “felony” and “felon” in online voter information for seven Florida counties. Each translation choice for the selected terms is noted during the parallel concordance review. In order to determine semantic shifts, the study searches for the relevant terms in available bilingual dictionaries and glossaries. Then, designations of semantic shifts are further supported by an examination of the terms’ usage in a parallel reference corpus – the bilingual laws of Puerto Rico, a Spanish-speaking US territory. The results of this study reveal a number of misrepresentative semantic shifts in both human- and machine-translated Spanish voter information. These shifts may impinge on individuals’ rights to political participation and civic engagement. While previous research has identified translation shifts that may distort Spanish-speaking individuals’ rights in the US legal system (Mellinger 2017), this study constitutes the first examination of this translation phenomenon in the context of LEP Spanish speakers’ political participation in the US.

**Keywords:** corpus-based translation studies, legal translation, political translation, voter rights, US elections, translation shifts, felon disenfranchisement

[6518] ***Is nostalgia what it used to be? The rhetorical functions of ‘nostalgia’ and ‘nostalgic’ over time.***

Anna Marchi (University of Bologna) and Charlotte Taylor (University of Sussex).

**Abstract.** Against an interdisciplinary backdrop of rising interest in the use of nostalgia in discourse, and its relationship to populism, this paper examines the lexical items ‘nostalgia’ and ‘nostalgic’ in order to uncover the rhetorical functions of these terms over an extended time period. As Marchi (2022) has shown, ‘nostalgia’ tends to convey a negative evaluative prosody (Morley & Partington 2009) in the UK press.

The aim of this paper is to reveal whether these evaluations are consistent over time and to identify both changing and shared targets of nostalgia across decades. Using the tools of corpus-assisted discourse analysis (e.g. Partington et al. 2013) we examine two continuous corpora of British English: Hansard and the Times which cover the time period from 1800-2022 and provide a measure of mainstream usage in public discourse. Although ‘nostalgia’ is first attested in the OED in 1756, and in the Times corpus in the 1870s, it does not gain currency in either corpus used here until the 1950s. Therefore, in this paper, we focus on the collocates from the 1950s onwards in order to understand how the term has evolved. This is done through two prisms. In the first, we identify the consistent, terminating, initiating and transient collocates, following McEnery & Baker (2016). In the second, we manually categorise collocates for each decade to enable comparison above the lexical level. In presenting our findings on the use of the term we focus on: a) who or what gets labeled as ‘nostalgic’ - for instance, the extent to which this can be a self- or other-describing label, b) the target of ‘nostalgia’ - for instance with reference to time periods ‘thirties’ is a collocate of nostalgia in 1950s/60s and 70s and then is replaced by later decades, in the 1980s and 1990s, c) the emotions with which ‘nostalgia’ is associated (e.g. wistfulness or melancholy), d) how ‘nostalgia’/being ‘nostalgic’ is evaluated and how this evaluation is used to (de)legitimise (e.g. 1950s key collocates include ‘agreeable’ whereas in the 2000s ‘soft-centred’ ‘dewy-eyed’ ‘sepia-tinted’ dominate), e) the use of metaphors in framing ‘nostalgia’ (e.g. LIQUID, SIGHT or HEAT). Our findings show that attitudes towards ‘nostalgia’ have changed considerably over the lifespan of its use as a mainstream term, while the targets have remained somewhat stable.

**Keywords:** nostalgia, collocation, corpus-assisted discourse studies, diachronic discourse analysis, media discourse, public discourse, emotion discourses, evaluative prosody

[6720] ***Key feature analysis and multi-dimensional analysis as ways to profile disciplinary variation in journal article writing.***

Paul Thompson (University of Birmingham) and Bethany Gray (Iowa State University).

**Abstract.** Systematic and functional descriptions of disciplinary variation in academic writing are a major focus in English for Academic Purposes (EAP) research. One approach to the study of disciplinary variation using corpus data is to examine a comprehensive range of features across disciplines through multi-dimensional (MD) analysis (for example, Gray 2015, Omidian et al. 2021). MD analysis is a powerful methodology that uncovers patterns of linguistic variation that can be interpreted functionally. However, MD analyses can be resource-intensive, and require a command of exploratory factor analysis.

Biber and Egbert (2018) have recently proposed a complementary and methodologically simpler procedure called 'key feature’ analysis, which relies on Cohen's d to identify (lexico)grammatical features that are markedly more/less frequent in a target corpus compared to a reference corpus, and which can result in 'profiles' of text varieties, which, in turn, can inform EAP pedagogy. Motivated by Egbert et al.'s (2020: 39) call to "[use] appropriate and minimally sufficient statistical methods", we explore what the simpler and more approachable key feature method can offer EAP researchers and practitioners.

Using a set of 120 functionally-motivated features, we apply both MD and key feature analysis to characterize disciplinary variation in the ELSE49 corpus (14,204 research articles, c. 72 million words, sampled from 49 academic journals and 10 disciplines). Variables include: traditional MD features from the Biber Tagcount programme (e.g., short passives, present tense); features characteristic of academic writing (e.g., existential there, exemplification markers); and newly-developed semantic categories (e.g., research nouns, analysis/interpretation verbs).

Our goal is not to determine a 'better' method for disciplinary variation research, but rather to compare what can be learned from a 5-dimension MD model and the identified key features (with each discipline compared to the remaining 9), and to consider in which context one approach may be more fitting than the other.

**Keywords:** key feature analysis, multi-dimensional analysis, disciplinary variation

[6940]***Spelling regularisation for Early Modern English texts: Advances in supervised and semi-supervised postediting.***

Andrew Hardie (Lancaster University).

**Abstract.** This paper presents an advance in spelling regularisation for Early Modern English made in the course of the AHRC-funded “Encyclopedia of Shakespeare’s Language” project (Culpeper and Archer 2020). A crucial aspect of this project was assorted exercises contrasting the language of Shakespeare’s plays to other datasets: Shakespeare’s verse, plays by other dramatists of the period, and general written English as exemplified by published books. Different corpus annotations were required for these exercises, all of which are however built on a foundation of spelling regularisation, the marking of words that have (by 21st century standards) non-standard spellings with standardised equivalents, to facilitate further annotation, querying, and statistical analysis.

The present state-of-the-art regularisation software is Baron’s VARD2 (https://ucrel.lancs.ac.uk/vard), which performs automatic regularisation based on prior “training” via manual regularisation. But automated regularisation has certain shortcomings. Its operations are not context-sensitive, so (for instance) cannot determine whether ‘ile’ should be normalised to ‘isle’ or ‘I’ll’. Nor can it recognise so-called “realword” errors (e.g. ‘doe’ for ‘do’, or ‘then’ for ‘than’). Manual postediting can address these issues, but requires greater effort than was practicable in our project.

We therefore implemented a novel program (‘Mortal Vard’) to support postediting by presenting a list of options for each word for the user to choose among at the stroke of a key, making it much faster to go through a whole text. Since the process is supervised, suggested changes can be less conservative than those of VARD2 itself, e.g. being generated by arbitrarily complex rules defined within the system’s resources. Finally, semi-supervised modes were implemented which apply ‘easy’ changes automatically while consulting the user for ‘hard’ words.

The presentation will include a demonstration of Mortal Vard.

Culpeper, J., & Archer, D. (2020). “Shakespeare’s language: Styles and meanings via the computer”. Language and Literature, 29(3), 191–202. <https://doi.org/10.1177/0963947020949436>

**Keywords:** Early Modern English, Spelling regularisation, Postediting, Corpus annotation, Software

[6984] ***Abusive name-calling: Representations around the term genocide during the covid pandemic.***

Yara Maria de Toledo Dias Romeiro (PUC-SP), Paula Tavares Pinto (São Paulo State University), Marcos Oliveira (Pontifical Catholic University of Sao Paulo) and Natalia Luri Arimori Ribeiro (Pontifical Catholic University of Sao Paulo).

**Abstract.** The Covid-19 pandemic has claimed over six million lives worldwide. In Brazil, the human tragedy was compounded by intense political polarization between the far right, and progressive, left-wing supporters, who objected to how former president Bolsonaro handled the pandemic. Twitter is a prime arena in which the conversation around the so-called COVID-19 genocide has taken place. Although the accusation of genocide was first leveled against opponents of the former president, his followers also adopted the term to accuse the current president of genocide. The question arises of whether it is possible to distinguish between the right-wing and the left-wing use of the term genocide. The goal of this paper is to map out the uses of genocide by supporters of each political wing and to determine if the patterns of usage are statistically different. To achieve this goal, a Lexical Multidimensional Analysis was carried out on a corpus of 150,000 tweets containing the terms ‘genocida’ (genocidist) and ‘genocídio’ (genocide).

A Lexical Multidimensional Analysis consists of a framework for detecting sets of correlated lexical units, which can be used to trace the underlying discourses around particular terms and/or groups (Berber Sardinha & Fitzsimmons-Dooley, in prep).

Methodologically, the approach consisted of scraping the tweets, tagging each post for part-of-speech, tabulating the lemma counts, and performing a factor analysis. The interpretation of the factors as dimensions was aided by careful consideration of large numbers of the highest-scoring tweets on each factor; these posts were read and interpreted for the underlying discourses indexed by each dimension. The dimensions were labeled to reflect the major discourses underlying the use of genocide.

Finally, the tweets were classified according to their right-wing or left-wing sources. In the presentation, the dimensions will be presented and illustrated and the major dimensional patterns distinguishing each political group will be discussed.

**Keywords:** Lexical Multidimensional Analysis, Infodemic Discourse, COVID-19 Pandemic, Corpus Linguistics, Discourse Analysis

[7112] ***Text-level measures of lexical dispersion: Robustness analysis.***

Lukas Sönning (University of Bamberg).

**Abstract.** The traditional approach to measuring lexical dispersion is to form corpus parts of equal size and then compare the occurrence rate of an item across these units. In recent methodological work, this strategy has met with criticism due to its ignorance to corpus structure (Biber et al. 2016; Burch et al. 2017; Egbert et al. 2020). Dispersion, it is argued, should be measured across linguistically meaningful units such as the individual text files constituting the corpus. A shift to the text as the unit of analysis is clearly desirable on linguistic grounds, and the ability of dispersion measures to handle unevenly-sized corpus parts has therefore received ample attention in the literature (e.g. Gries 2008, 2020). The question of how existing metrics perform on these data layouts has only been partly addressed, however. The aim of this paper is to shed light on the statistical

robustness of seven text-level dispersion measures (D, D2, Sadj, DP, DA, DKL and text dispersion; see Gries 2020; Egbert & Biber 2019; Egbert et al. 2020). The focus is on two aspects: (i) whether they are systematically distorted by the number of texts in the corpus (as is the case for Juilland’s D; Biber et al. 2016); and (ii) whether they are (overly) sensitive to data situations that can arise when texts differ (considerably) in length. We use hypothetical data scenarios to identify weak spots in existing measures, and then propose modifications to DP- and DA-related indexes to effect more robust estimators. Along with the other measures, these are then evaluated against actual corpus data drawn from the BNC. We observe that adapted DP- and DA-variants perform at least as well as their original versions. Our permutation-based simulation study also demonstrates that Carroll’s D2 shows the same weakness as Juilland’s D, i.e. a noticeable sensitivity to the number of units that enter the analysis.

**References**

Biber, Douglas, Randi Reppen, Erin Schnur & Romy Ghanem. 2016. On the (non)utility of Juilland’s D to measure lexical dispersion in large corpora. International Journal of Corpus Linguistics 21(4), 439–464.

Burch, Brent, Jesse Egbert & Douglas Biber. 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, Jesse & Douglas Biber. 2019. Incorporating text dispersion into keyword analysis. Corpora 14(1), 77–104.

Egbert, Jesse, Brent Burch & Douglas Biber. 2020. Lexical dispersion and corpus design. International Journal of Corpus Linguistics 25(1), 89–115.

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

Gries, Stefan Th. 2020. Analyzing dispersion. In Magali Paquot & Stefan Th. Gries (eds.), A practical handbook of corpus linguistics. New York: Springer, 99–118.

**Keywords:** dispersion measures, lexical dispersion, corpus design, text-level analysis, robustness

[7148] ***Coding Discourse marker functions in corpora of spoken English.***

Nuha Alharbi (Lancaster University).

**Abstract.** This paper presents a corpus-discourse analysis of the functions of discourse markers (DMs) in two corpora: a corpus of Arabic children learning English as a second language (Ar-children) and a customized children corpus from the spoken BNC2014.

Building on previous studies analyzing DMs and their functions on native and learner corpora, this paper has a two-fold aim: First, it provides an empirical analysis of DM functions in the peer-talk of child second language learners. Second, it will discuss challenges in analyzing DMs according to a proposed framework of functions.

The analysis of DMs is beneficial to expand our understanding of language learners’ pragmatic competence and as such their second language competence. Using a mixed-method approach, the present study attempts to offer answers about the frequencies and functions of English DMs in the speech of Arabic-speaking children in comparison to their usage in age-matched English-speaking children. However, the qualitative coding of DM functions is not straightforward due to their multifunctionality (Crible & Cuenca, 2017: 155).

Previous research has tended to code DMs according to two main functional domains: textual and interpersonal (cf. Müller, 2005; Aijmer, 2013; Brinton, 2006, 2017). In the textual domain, DMs mark the structuring of discourse and the interpersonal domain captures functions that signal the speaker’s attitude and the management of social exchange (Brinton, 2006: 310). In my application of a coding scheme for documenting DM functions, I show that the distinction between the two domains is not a clear-cut, which has implications for how we count and analyze their use. By offering examples of DMs taken from authentic learner language, this work demonstrates that DMs can serve multiple, simultaneous functions and this needs to be taken into account when investigating language competence.

**References**

Aijmer, K. (2013). Understanding pragmatic markers in English. Edinburgh University Press.

Brinton, L. J. (2006). Pathways in the development of pragmatic markers in English. In A. van Kemenade & B. Los (Eds.), The handbook of the history of English (pp. 307–334). Blackwell Pub.

Brinton, L. J. (2017). The evolution of pragmatic markers in English: Pathways of change. Cambridge University Press.

Crible, L., & Cuenca, M.-J. (2017). Discourse Markers in Speech: Distinctive Features and Corpus Annotation. Dialogue & Discourse, 8(2), 149–166. <https://doi.org/10.5087/dad.2017.207>

Müller, S. (2005). Discourse markers in native and non-native English discourse. John Benjamins Pub.

**Keywords:** corpus linguistics, Discourse markers, Pragmatic competence, Child second language acquisition

[7162] ***Down-sampling from hierarchically structured corpus data: The case of third-person verb inflection in Early Modern English.***

Lukas Sönning (University of Bamberg) and Ole Schützler (University of Leipzig).

**Abstract.** When resource constraints preclude the analysis of all tokens returned by a corpus query, researchers use ‘down-sampling’ and (randomly) select a subset of cases. It has been shown that drawing a simple random sample from the full body of corpus hits may not be the most efficient approach (Sönning & Krug 2021). Thus, if interest centers on predictors measured at the level of the speaker (e.g. age and gender), information about the source of a token (i.e. speaker or author) can be used to maximize the information in a sub-sample.

This paper extends the evaluation of down-sampling schemes to settings where tokens are clustered not only by source but also on linguistic grounds. Our case study deals with the development of third-person verb inflection from -(e)th to -(e)s during the Early Modern English period. We rely on data from Jenset & McGillivray (2017), which are drawn from the PPCEME corpus (Kroch et al. 2010). The 12,337 corpus hits are clustered by source (150 authors) and verb (1,009 lemmas, excluding BE). Our focus is on five predictors (cf. Kytö 1993; Nevalainen & Raumolin-Brunberg 2003; Gries & Hilpert 2010): three author-level variables (time, gender, genre), and two verb-level variables (frequency, final sibilant).

Assuming our budgetary limitations allow us to analyze 2,000 tokens, we compare two strategies for selecting a subset of this size: (i) simple random down-sampling, where each hit has the same probability of being selected; and (ii) structured random down-sampling, where this probability is inversely proportional to the author- and verb-specific token count. We form 500 subsets using each scheme and compare estimates based on mixed-effects logistic regression to a reference model fit to all 12,337 cases. We observe that structured down-sampling outperforms simple down-sampling in terms of average down-sampling error (absolute deviation from benchmark), bias (systematic deviation from benchmark) and precision (statistical uncertainty relative to benchmark).

**References**

Gries, Stefan Th. & Martin Hilpert. 2010. Modeling diachronic change in the third person singular: A multifactorial, verb- and author-specific exploratory approach. English Language and Linguistics 14(3), 293–320.

Jenset, Gard B. & Barbara McGillivray. 2017. Quantitative historical linguistics: A corpus framework. Oxford: Oxford University Press.

Kroch, Anthony, Beatrice Santorini & Ariel Diertani. 2010. Penn Parsed Corpus of Modern British English. <http://www.ling.upenn.edu/hist-corpora/PPCMBE-RELEASE-1/index.html>

Kytö, Merja. 1993. Third-person singular verb inflection in early British and American English. Language Variation and Change 5(2), 113–39.

Nevalainen, Terttu & Helena Raumolin-Brunberg. 2003. Historical sociolinguistics: Language change in Tudor and Stuart England. London: Pearson Education.

Sönning, Lukas & Manfred Krug. 2022. Comparing study designs and down-sampling strategies in corpus analysis: The importance of speaker metadata in the BNCs of 1994 and 2014. In Ole Schützler & Julia Schlüter (eds.), Data and methods in corpus linguistics: Comparative approaches. Cambridge: Cambridge University Press, 127–160.

**Keywords:** down-sampling, corpus design, hierarchical data structures, efficiency, bias

[7223] ***Linguistic characteristics of pseudo-science: A corpus-based study.***

Deise Prina Dutra (Minas Gerais Federal University), Ana Bocorny (Rio Grande do Sul Federal University), Maria Claudia Delfino (Pontifical Catholic University of São Paulo), Paula Tavares Pinto (São Paulo State University), Simone Sarmento (Rio Grande do Sul Federal University) and Tony Berber Sardinha (Pontifical Catholic University of São Paulo).

**Abstract.** Although the COVID-19 infodemic is normally associated with social media, academic discourse has been a mainstay of denialist groups around the world as they propagate disinformation about the pandemic. These movements often cite pseudo-science as the basis for their claims, arguing against recommendations based on verifiable scientific evidence put forth by governing bodies and health organizations. Few studies, thus far, have addressed the infodemic in academic discourse. One study is Pflugfelder (2022), which looked at rhetorical features linked to characteristics such as ethos, logos, style, and evidence in 20 fraudulent articles using content analysis.

This paper presents a corpus-based study of 212 pseudo-scientific articles, comprising 1,882,344 tokens, which seeks to identify discourse features that characterize pseudo-scientific articles. Our analytical framework is Lexical Multidimensional (LMD) Analysis (Berber Sardinha, 2019, 2021; Berber Sardinha & Fitzsimmons-Dooley, in prep; Fitzsimmons-Dooley, 2019, forthcoming; Clarke et al. 2022), which enables the identification of discourses through the application of multivariate statistical techniques.

The procedures involved collecting a corpus of COVID-19 pseudo-science articles published in English, tagging it for POS and lemmatizing the vocabulary. The lemma counts were computed and entered in a factor analysis, which determined the major factors, each comprising a set of correlated lexical features.

Using Sketch Engine, we extracted keywords in comparison to a separate English reference corpus consisting of non-academic texts, yielding keywords like hydroxychloroquine (N=4,704), sars-cov-2 (N=3,314), azithromycin (N=1,513) and chloroquine (N=1,258). Careful analysis of concordances of these words revealed key strategies employed by pseudo-science authors.

For instance, with hydroxychloroquine as the subject, frequent verb collocates include ‘reduce’, ‘inhibit’ and ‘decrease’, as in ‘low dose of hydroxychloroquine reduces fatality of critically ill patients’, which promotes a treatment whose effectiveness has been widely denied by science.

**Keywords:** coronavirus infodemic, pseudo-science, academic articles, linguistic patterns

[7236] ***Progress, innovations, and challenges to date in the compilation of LANA, a corpus of American English.***

Elizabeth Hanks (Northern Arizona University), Tony McEnery (Lancaster University), Vaclav Brezina (Lancaster University), Jesse Egbert (Northern Arizona University), Tove Larsson (Northern Arizona University), Doug Biber (Northern Arizona University), Randi Reppen (Northern Arizona University), Paul Baker (Lancaster University), Gavin Brookes (Lancaster University), Raffaella Bottini (Lancaster University) and Isobelle Clarke (Lancaster University).

**Abstract.** This presentation introduces the design and collection of a new corpus of American English: the Lancaster-Northern Arizona Corpus of American English (LANA). The goal of the project is to compile a publicly available, large-scale corpus of both written (LANA-CAWE) and spoken (LANA-CASE) American English which can serve as a comparable counterpart to the BNC2014 (Brezina et al., 2021; Love et al., 2017).

We will discuss key principles of LANA-CAWE and LANA-CASE design, which builds on strategies established in the creation of the BNC2014 but also implements theoretical and technical innovations that have emerged in recent years. With the goal of informing future corpus-construction projects, we will discuss the following:

1. Design: we design the corpus by following methods proposed in Egbert, Biber, and Gray (2022), including describing the domain, operationalizing the domain, planning the sample, and evaluation;

2. Sampling: we conduct several rounds of data collection. For the written section, data is sampled from a wide range of sources including academic journals, academic books, fiction books, newspapers, magazines, and web registers such as blogposts and social media posts. For the spoken portion, we utilize social media, public outreach efforts, and market research panels to recruit participants from diverse regions, race/ethnicities, ages, and settings (urban/suburban or rural);

3. Data collection: we collect written texts that are available in the public domain as well as those made available to us by publishers and participants. We leverage PPSR (public participation in scientific research; Shirk et al., 2012) by collecting spoken data through online questionnaires that allow participants to contribute remotely;

4. Progress to date: we summarize the progress made, including a review of the representativeness of both written and spoken strata and plans to complete and disseminate the corpus.

**References**

Brezina, V., Hawtin, A., & McEnery, T. (2021). The Written British National Corpus 2014–design and comparability. Text & Talk, 41(5-6), 595-615. <https://doi.org/10.1515/text-2020-0052>

Egbert, J., Biber, D., & Gray, B. (2022). Designing and evaluating language corpora: A practical framework for corpus representativeness. Cambridge University Press.

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>

Shirk, J. L., Ballard, H. L., Wilderman, C. C., Phillips, T., Wiggins, A., Jordan, R., McCallie, E., Minarchek, M., Lewenstein, B. V., Krasny, M. E., & Bonney, R. (2012). Public participation in scientific research: a framework for deliberate design. Ecology and Society, 17(2). <https://www.jstor.org/stable/26269051>

**Keywords:** corpus compilation, spoken English, written English, methodology

[7252] ***Exploring cross-linguistic representations of Web registers with a deep multilingual model.***

Aki-Juhani Kyröläinen (University of Turku), Filip Ginter (University of Turku) and Veronika Laippala (University of Turku).

**Abstract.** Register is considered one of the best predictors associated with linguistic variation [1] and contemporary language models can be utilized to accurately model this type of variation within languages [5]. However, linguistic studies have offered evidence that information pertaining to register variation can be grounded in universal tendencies [see 2, 7]. Rather than modeling each language with a separate model, multilingual models have recently emerged that can accurately capture register variation in a cross-linguistic setting by learning a shared representation [see 8]. However, this opens up the question of the linguistic nature of this shared representation. In this study, we specifically investigate this question and present results based on a model trained on six Web registers across three languages (English, French and Finnish). To the best of our knowledge, this is the first study to embark on this type of large-scale investigation into universal tendencies associated with register variation.

Comparable multilingual web corpora were extracted from Register Oscar [6] to fine-tune a multilingual XLM-R model to predict six registers across the three languages [3]. The model accurately discriminated the cross-linguistic register classes (F1-score of 80%). To analyze the shared representation of the model, for each of the 13 layers, we extracted the semantic vectors associated with 100 keywords from each register and language using SACX [9]. Finally, these vectors were clustered separately for each layer using model-based clustering with mixtures from von Mises-Fisher distributions [4].

The results demonstrated that the register-specific representation associated with the keywords fully emerged at the deepest layer, i.e., the clusters were formed based on the registers instead of the languages. Additionally, a within-cluster analysis showed that the keywords could be linked to functional properties of the registers such as semantic verb categories and tense. The findings are discussed in relation to cross-linguistic register variation.

**References**

[1] Douglas Biber. “Register as a predictor of linguistic variation.” In: Corpus linguistics and linguistic theory 8.1 (2012), pp. 9–37. DOI: 10.1515/cllt-2012-0002.

[2] Douglas Biber. “Using multi-dimensional analysis to explore cross-linguistic universals of register variation.” In: Languages in Contrast 14.1 (2014), pp. 7–34. DOI: 10.1075/bct.87.02bib.

[3] Alexis Conneau et al. “Unsupervised cross-lingual representation learning at scale.” In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020, pp. 8440–8451. DOI: 10.18653/v1/2020.acl- main.747.

[4] Kurt Hornik and Bettina Gr ̈un. “movMF: an R package for fitting mixtures of von Mises-Fisher distributions.” In: Journal of Statistical Software 58.10 (2014), pp. 1–31.

[5] Veronika Laippala et al. “Register identification from the unrestricted open Web using the Corpus of Online Registers of English.” In: Language Resources and Evaluation (2022), pp. 1–35. DOI: 10.1007/s10579-022-09624-1.

[6] Veronika Laippala et al. “Towards better structured and less noisy Web data: Oscar with Register annotations.” In: Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022). 2022, pp. 215–221.

[7] Haipeng Li, Jonathan Dunn, and Andrea Nini. “Register variation remains stable across 60 languages.” In: Corpus Linguistics and Linguistic Theory (2022). DOI: 10.1515/cllt-2021-0090.

[8] Liina Repo et al. “Beyond the English Web: Zero-Shot Cross-Lingual and Lightweight Monolingual Classification of Registers.” In: Proceedings of the EACL 2021 Student Research Workshop. 2021.

[9] Samuel Rönnqvist et al. “Explaining classes through stable word attributions.” In: Findings of the Association for Computational Linguistics: ACL 2022. 2022, pp. 1063–1074.

**Keywords:** register variation, register universals, explainable deep language models, neural networks, Black-box models

[7315] ***Corpus-based Bilingual Terminology Extraction using One-Click Terms.***

Miloš Jakubíček (Lexical Computing), Ondrej Matuska (Lexical Computing) and Marek Blahuš (Lexical Computing).

**Abstract.**

In this paper we describe recent development of the One-Click Terms system which automatically extracts terminology using the corpus-based contrastive methodology built in the Sketch Engine corpus management system.

The terminology extraction as implemented in Sketch Engine generates term candidates using a language-specific term grammar (terms are typically noun phrases) and scores these candidates by comparing the normalized frequencies with a large reference corpus [1].

The system has been later extended to bilingual terminology extraction from aligned documents [2] based on co-occurrences in aligned segments being ranked using the logDice association score [3].

The development that we present in this paper covers implementation of the aforementioned technology in a dedicated standalone tool outside of Sketch Engine which uses Sketch Engine to create corpora under the hood but spares the user from being aware of the corpus-related machinery. This tool integrates a Church-Gale-inspired automatic aligner [4] so that it works also with unaligned bilingual texts which users can upload and immediately see the extracted bilingual terminology. A large number of input formats is supported (TMX, XLIFFv2, PDF, DOC, DOCX, HTML, TXT) and the extracted terms can be downloaded in the form of a CSV, XLS or TBX file.

The language coverage of One-Click Terms has also been extended to 24 languages. New languages were added and existing ones were improved by devising term grammars based on the IATE terminological database that is being developed by the Translation Centre for the Bodies of the European Union [5] focusing on high coverage, yielding for English 74 % of all IATE terms.

**References**

[1] JAKUBÍČEK, Miloš, et al. Finding terms in corpora for many languages with the Sketch Engine. In: Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics. 2014. p. 53-56.

[2] KOVÁŘ, Vojtěch; BAISA, Vít; JAKUBÍČEK, Miloš. Sketch engine for bilingual lexicography. International Journal of Lexicography, 2016, 29.3: 339-352.

[3] RYCHLÝ, Pavel. A Lexicographer-Friendly Association Score. In: RASLAN. 2008. p. 6-9.

[4] GALE, William A., CHURCH, Ken. A program for aligning sentences in bilingual corpora. Computational linguistics, 1994, 19.1: 75-102.

[5] JOHNSON, Ian; MACPHAIL, Alastair. IATE-Inter-Agency Terminology Exchange: development of a single central terminology database for the institutions and agencies of the European Union. In: Workshop on Terminology resources and computation. 2000.

**Keywords:** bilingual term extraction, One-Click Terms, Sketch Engine

[7324] ***‘Roe was egregiously wrong from the start’: A Contrastive Analysis of the Attitudinal Evaluation of Abortion Case Law by SCOTUS Opinion Writers.***

Jamie McKeown (The Hong Kong Polytechnic University) and Mary Ye (The Hong Kong Polytechnic University).

**Abstract.** Whilst evaluation has been well investigated in judicial discourse (e.g. see, Mazzi, 2015) there is little language scholarship on the interplay of evaluation and judicial political leaning.

In this presentation, we investigate a context where judicial political leaning has had a demonstrable impact, i.e., the abortion law of the U.S. Across 50 years of case law, SCOTUS (see Ziegler, 2016: xxix-xxx) has oscillated between liberal and conservative directions in its adjudication of abortion rights. Using the appraisal framework (Martin and White, 2005) we contrast the use of attitude by liberal and conservative judicial opinion writers in their evaluation of SCOTUS abortion case law.

Here, we focus on majority and dissent opinions as they provide a clear line of divide within any given case (McKeown, 2022). In doing so, we use 4 corpora (339,310 words in total): a conservative majority corpus which we compare to a liberal majority corpus; and, a conservative dissent corpus which we compare to a liberal dissent corpus. Each corpus contains 10 opinions with 2 opinions taken from each of the 5 decades in which Roe was legally binding. We extracted all explicit references to the current case and past abortion case law in the respective corpora (3,291 concordance lines).

Taking into account opinion type (i.e., majority or dissent) as well as the evaluative target (i.e., current case or previous case), we annotated the concordance lines for the use of attitude in order to answer the following question: did the use of appraisal to evaluate SCOTUS abortion case law differ according to political leaning?

Our findings show that whether writing for the majority or in dissent liberal opinion writers used significantly more positive attitude to essentially extoll the virtues of this area of law. Conversely, conservative opinion writers used a significantly greater amount of negative attitude to attack and undermine this area of law thus ‘setting the stage for its [eventual] overruling’ (Ziegler, 2020: 8). In the presentation, we will discuss other salient findings as well as the underlying ideation of individual subcategories of attitude.

**Keywords:** Corpus Assisted Discourse Analysis, Judicial Discourse, Appraisal, Political Leaning, Roe v Wade

[7346] ***Indicating engagement in online workplace meetings: The role of backchannelling head nods.***

Anne O'Keeffe (Mary Immaculate College), Dawn Knight (Cardiff University), Christopher Fitzgerald (Mary Immaculate College), Justin McNamara (Mary immaculate College), Geraldine Mark (Cardiff University), Svenja Adolphs (Nottingham University), Sandrine Peraldi (University College Dublin), Tania Fahey Palma (University of Aberdeen), Fiona Farr (University of Limerick) and Benjamin Cowan (University College Dublin).

**Abstract.** There is a need, particularly in a post-COVID world that has witnessed a ‘digital pivot’, to better understand what comprises effective communication online, and to better understand virtual communication itself. Research being carried out on the ‘Interactional Variation Online’ (IVO) project aims to address this need by undertaking corpus-based multi-modal analyses of virtual workplace communication to gain depth of insight into the potential barriers and carriers of effective dyadic and multi-party talk in this context.

Within the context of this project, this presentation explores the dynamics of indicators of engagement in multi-party communication online, as evidenced by a small-scale but in-depth corpus-based multi-modal study by investigating how verbal and non-verbal behaviour act individually and combine to function as backchannels in virtual meetings. This is achieved through a multimodal analysis of spoken backchannels and nods that function as backchannels, using ELAN, in a 40-minute sub-corpus of the beginnings and ends of four of the virtual meeting recordings.

Findings reveal that participants in this setting:

1. use head nods over spoken backchannels to indicate listenership,

2. minimise the cooccurrence of head nods and spoken backchannels,

3. minimise the use of continuing head nods and spoken backchannels because their valency in terms of pragmatic force appears to have changed in this medium,

4. favour displays of convergence through both head nods and spoken backchannels, and

5. accommodate towards each other’s convergent response behaviour in clusters, especially at ends of meetings.

Overall, backchanneling head nods appear to have taken on greater importance as minimal non-verbal backchannels in online meetings, but this is also indicative of changes in the nature of turn-taking, turn yielding and turn length that merit further investigation, which, we argue, multi-modal corpora can facilitate.

**Keywords:** backchannelling, multi-modality, virtual communication, response tokens

[7356] ***Developing a youth talk corpus for Turkish: The Corpus of Turkish Youth Language (CoTY) and the use of vocatives as interactional markers.***

Esranur Efeoğlu-Özcan (Gazi University) and Hale Işık-Güler (Middle East Technical University).

**Abstract.** This talk presents the design criteria and structural characteristics of the Corpus of Turkish Youth Language (CoTY), which is the first youth talk corpus compiled for contemporary spoken Turkish, and provides an overview of pragmatic functions of vocatives in this corpus. The CoTY is a 168,748-word corpus within the single register of informal conversation exclusively among friends. Designed to offer a maximally representative sample of Turkish youth talk, the CoTY includes naturally occurring and spontaneous interactional data among 123 unique speakers (62 females and 61 males) between the ages of 14-18 from various socio-economic backgrounds in Turkey. The corpus was constructed using the multilayer transcription and corpus construction software EXMARaLDA, the tools of Partitur-Editor, COMA, and EXAKT were utilized as corpus building, management, query and analysis tools. The construction of this corpus enables the previously unattained research area of contemporary spoken Turkish used in dyadic and multi-party interaction among young speakers of Turkish. After describing the architecture of the corpus, this talk will focus on the nominal vocatives which are used as interactional markers among Turkish speaking youth. The overview of nominal vocatives in the CoTY with regard to addressers, addressees, semantic categories, pragmatic extension, address shifts, register-specific uses indicate that young speakers of Turkish use vocatives to attend to both organizational and interpersonal needs they experience in interaction with their friends. The results corroborate the perspective that the repertoire of

vocatives is extensively used to project and enhance the intimate level of relationship the speakers have by attending to face concerns and creating playful and humorous tone in interaction. It is expected that the results of this study will provide baseline data for further studies on contemporary spoken Turkish and cross-linguistic youth language studies.

**Keywords:** youth language, spoken discourse, vocatives, Turkish

[7374] ***Cross-validation across multivariate analytical systems: A statistical method for the CEFR-level estimation and automated assessment of the syntactic/lexical complexity of English-L2 learner texts.***

Andrea Agnes Remenyi (Pazmany Peter Catholic University) and Karlygash Adamova (Pazmany Peter Catholic University).

**Abstract.** A recurring language assessment question is how to conceptualise language proficiency levels according to the Common European Framework of Reference for Languages (CEFR) in quantifiable features of English grammar and vocabulary. In other words, what are the characteristics of a certain CEFR level, in terms of not the language skills but its syntactic and lexical accuracy and complexity of both? A limited, though not necessarily minimal, set of those quantifiable characteristics could inform the teaching-learning process and also facilitate spoken and written L2-learner text assessment. Moreover, such a set of variables could work towards the automated assessment of learner texts.

We set out to detect the systematic patterns of syntactic and lexical characteristics of a written learner corpus and their match to CEFR level expectations (in our case, it is the B2+ CEFR level). The corpus (414 texts, 89,838 running words) contains cca. 200 word formal letters and narratives written on the basis of prompts at a high-stakes language examination targeting the B2+ level.

The multivariate research design examines the influence of numerous variables in the patterning of syntactic and lexical complexity in those texts, including automatically and manually detectable ones.

We have been working with the Biber-tagger/MAT (Nini 2019), the L2SCA (Lu 2017), the CVLA (Uchida & Negishi 2018), Lextutor-based measures (Cobb 2022), and some of our own automated and manual measures, including also the scores of human raters (Reményi-Velner 2022, Adamova 2022).

All these variables were statistical analysed through correlation and factor analysis to unearth the patterns that are the most salient across these analytical systems and thus cross-validate each other. For example, the Lextutor-based K2up/K1 variable and the two vocabulary variables in the CVLA corroborate each other. Similarly, sentence/clause length (L2SCA) and demonstratives (MAT) align in the same latent variable.

**Keywords:** English-L2 learner corpus, syntactic/lexical complexity analysis, CEFR-level estimation of learner texts, automatic language assessment

[7381] ***From Argentina to Zimbabwe: Exploring the global appeal of the International Baccalaureate 1977-2019.***

Saira Fitzgerald (Lancaster University).

**Abstract.** This study presents the third stage of a research project examining perceptions of the International Baccalaureate (IB) over five decades to better understand its growing influence on education systems around the world. As part of the global education industry (Verger et al., 2016), the IB plays an important role in shaping national public education discourse. Despite its popularity across diverse social and political contexts, understanding of this educational institution is limited, relying predominantly on information produced by the IB or its affiliates (Dugonjić-Rodwin, 2021). The IB thus tends to be presented uncritically in laudatory terms that get repeated across contexts, such that they eventually obscure underlying contradictions and inaccuracies, e.g., the contested histories of the IB that shift pragmatic (university admission) emphasis to idealistic (world peace) endeavours.

To explore how the IB is perceived in different countries, a 27 million word specialized corpus of global press articles was created as an unsolicited window into public opinion (Mautner, 2008) and included all sub-registers such as letters to the editor, opinion pieces, obituaries, news and sports. Collocation, cluster and concordance analyses were conducted in three stages: (1) corpus as a whole; (2) change over time; and (3) country by country comparison, the focus here.

Results showed that the IB is typically found in the context of debates about education systems, where it is compared to national curricula through three main frames: (i) as a model for education reform (“falling standards” discourse); (ii) as one of many possibilities or alternative (discourse of choice); and (iii) evaluative statements (discourse of superiority). Trends across time revealed a notable shift from concrete curriculum-related lexis to recurring value-laden terms, suggesting one way the IB appears polysemous, allowing it to mean different things in different contexts according to stakeholder need and thus appeal to countries across the ideological spectrum.

**Keywords:** corpus linguistics, CADS, International Baccalaureate, IB, global education, national education, media

[7395] ***A Corpus-based Study of Psycholinguistic Signs of Alzheimer’s Disease in Written Language.***

Xufei An (School of Foreign Studies, Xi'an Jiaotong University) and Yue Jiang (School of Foreign Studies, Xi'an Jiaotong University).

**Abstract.** While extensive studies have been conducted on spoken language impairment in patients with Alzheimer’s disease (AD), little is known concerning their use of psycholinguistic signs in written language. To address this niche, this study collects as a corpus of 26 fictions by Iris Murdoch (the AD group) and 19 by P. D. James (the healthy control)and annotates with the Stanford CoreNLP (4.4.0) parser, and investigates whether their use of dashes and filled pauses correlates with their age or the progression of AD.

The results show that, compared with P. D. James, (i) there are significant positive correlations between the frequency of dashes (r = 0.859, p < 0.001), filled pauses (r = 0.624, p < 0.01), and I.M.’s age, with the frequency of dashes rising most significantly in her last novels; (ii) the frequency of dashes and filled pauses in I.M.’s works also significantly correlates with one another (r = 0.743, p < 0.001); and (iii) the increasing use of dashes and filled pauses in I.M.’s fictions interrupts the information flow, hinders the textual cohesion and coherence, and thus possibly attests to her cognitive abilities impaired by AD. These findings suggest that dashes and filled pauses may serve as important psycholinguistic signs of cognitive impairment of AD patients, and should be considered in the design of rating scales and clinical assessments for AD.

**Keywords:** Corpus, Psycholinguistic signs, Dashes, Filled pauses, Written language, Alzheimer’s disease

[7426] ***Corpus Approach to Mental Health Discourse on Instagram: How Influencers Communicate Vulnerabilities.***

Aleksandra Sevastianova (University of Edinburgh).

**Abstract.** Mental health dialogue is becoming gradually destigmatized within society, with online communication being an important contributor. On some online platforms, such as specialized forums and groups, mental health has always been a central topic. However, other platforms, such as Instagram, have started to address mental health only recently, which signifies a shift in social attitudes to mental struggle. Lifestyle Insta-influencers, who used to commercialize a flawless self-image, have now shifted towards authenticity, where sharing sensitive issues is no longer a sign of weakness (Foos, 2021). This talk aims to explore the ways in which influencers articulate mental struggle to align the sensitive discourse with their successful and knowledgeable online persona.

This research presents intermediate results of the ongoing 2nd year PhD project. Employing corpus-assisted discourse analysis, it builds upon previous online health discourse studies (Hunt & Harvey, 2015; Feuston & Piper, 2019) and is informed by the “small stories” framework (Georgakopoulou, 2007) as well as discourse-centred online ethnography (Androutsopouos, 2008).

For the research, we manually selected 6 most popular lifestyle bloggers known to post about such mental health conditions as anxiety, depression, and eating disorder. Ethnographic method was used to characterize influencers’ online personas. Using Apify tools, posts were scraped and turned into a #LancsBox corpus. Keyword and concordances analysis was undertaken, followed by narrative analysis of selected mental health posts.

The data analysis demonstrated that influencers tend to communicate mental health in the ways that align with their self-mediated image, presenting vulnerabilities as integral aspects of a successful individual. Among other ways, it may be achieved by using two discursive strategies - authentication and aestheticization of mental health, which allows bloggers to keep the balance between sincerity and attractiveness. The positive way in which influencers communicate vulnerabilities might have an impact on the social perception of mental health struggles.

**References**

Androutsopoulos, J. (2008) Potentials and Limitations of Discourse-Centred Online Ethnography.

Feuston, J.L., & Piper, A.M. (2019) Everyday Experiences: Small Stories and Mental Illness on Instagram. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.

Foos, A. (2021) Adversity as Authenticity: How Influencers Use Personal Struggles to Build their Brands. The BRC Academy Journal of Business, 11(1). pp.117-141.

Georgakopoulou, A. (2007) Small Stories, Interaction and Identities. Studies in Narrative. No.8. Amsterdam: John Benjamins.

Hunt, D., & Harvey, K. (2015) Health communication and corpus linguistics: using corpus tools to analyse eating disorder discourse online. In Corpora and Discourse Studies (pp. 134-154). Palgrave Macmillan, London.

**Keywords:** corpus-assisted discourse analysis, online health communication, mental health discourse, Instagram discourse

[7455] ***The Discourses of Climate Change Denialism across Conspiracy and Pseudoscience Websites.***

Isobelle Clarke (Lancaster University).

**Abstract.** This talk will describe the common discourses of climate change denialism in a corpus of texts mentioning climate change and global warming from 186 websites and blogs known to promote pseudoscience and conspiracy. To identify the discourses, this corpus of texts from pseudoscience and conspiracy websites mentioning climate change and global warming was compared to the English Web 2020 corpus (enTenTen20) for keywords. This list of keywords was then analysed using Keyword Co-occurrence Analysis (KCA) (Clarke et al., 2021). KCA is an approach for identifying the most common patterns of co-occurring keywords across the texts of a corpus.

In the present analysis, each text in the corpus of texts from pseudoscience and conspiracy websites was automatically analysed for the presence/absence of each of the keywords using a computer programme. This information was then recorded in a data matrix. This data matrix was then subjected to Multiple Correspondence Analysis, which groups keywords into dimensions based on their frequent co-occurrence across the texts. These dimensions of co-occurring keywords were then interpreted for the discourses to which they point.

The analysis revealed that all the tactics and arguments known to have been developed and promulgated by conservative think-tanks (funded by the fossil fuel industry) have permeated online blogs and websites promoting pseudoscience and conspiracy. Additionally, the analysis shows that while outright denial of climate change is common on these websites, texts are more likely to focus on downplaying the seriousness of climate change. It is argued that downplaying the seriousness of climate change is a subtler and more believable strategy in comparison to outright denial, which is harder to believe in the face of growing scientific evidence of anthropogenic climate change.

Clarke, I., McEnery, T. and Brookes, G. (2021) Multiple Correspondence Analysis, newspaper discourse and subregister: A case study of discourses of Islam in the British press. Register Studies 3(1): 144-171.

**Keywords:** Multiple Correspondence Analysis, Keyword Co-occurrence Analysis, Keywords, Climate change, Conspiracy theories, Misinformation

[7545] ***Corpus linguistics for audiovisual translation revisited: Netflix as a multimedia corpus.***

Judyta Mężyk (Tischner European University in Cracow, University of Silesia in Katowice, Paris-East Créteil University).

**Abstract.** Both audiovisual translation (AVT) research and corpus studies have been a topic of interest for an exponential number of scholars in recent years. Therefore, it is no surprise that more and more effort is put towards using corpus methods for AVT purposes. However, over ten years after the introduction of the Veiga Corpus, a promising multimedia corpus (e.g., Sotelo Dios 2015), and exactly ten years after the special issue of Perspectives journal devoted to AVT from the perspective of corpus linguistics, the number of proper corpora that could be used for such examinations is still limited. The reason for this is that AVT should be researched by not only considering the textual layer since more factors are taken into account in such a translation (Baldry and Taylor 2004, Gambier 2014); thus, for researching AVT, multimedia corpora should be used, but building them requires a great amount of time and effort. This paper suggests a solution to this problem: the use of Netflix as a multimedia corpus not only of subtitles but also of audio and visual information. Firstly, Language Reactor, a Google Chrome extension thanks to which this solution is possible, is discussed with a focus on its feature of subtitles exportation and what such an exported file contains. Then, a study on situational multi-word units (pragmatemes) is presented as an example of an examination of such a corpus with the use of corpus processing software (namely Unitex and AntConc). Finally, future prospects of this method are discussed since it is argued that with such a massive corpus as offered by the Netflix library, the possibilities for AVT research using corpora are bigger than ever.

References to the abstract available on Google Drive: <https://docs.google.com/document/d/1MYyo9CHZ24cVreR06GMa5Cwers3CyuqjE37vDXQsNlQ/edit?usp=sharing>

**Keywords:** audiovisual translation, corpus linguistics, subtitling, pragmatemes

[7864] ***Robustness of multi-dimensional register analysis with low-quality data: The case of Scottish Enlightenment writing.***

Aatu Liimatta (University of Helsinki), Yann Ryan (University of Helsinki), Mikko Tolonen (University of Helsinki) and Tanja Säily (University of Helsinki).

**Abstract.** While large-scale textual datasets are extremely powerful, corpus-linguistic analyses making use of them are often hampered by the low quality of the textual data. While the problem can affect many datasets, such as social media (Eisenstein 2013), it is, arguably, often at its worst in large-scale historical datasets such as Eighteenth Century Collections Online (ECCO), whose machine-readable versions have been produced using optical character recognition (OCR) and are of questionable quality (Hill & Hengchen 2019). The problem is particularly salient for approaches which require the correct identification of longer, more complex linguistic constructions. We however hypothesize that analyses which aggregate a large spread of items and constructions in the analysis pipeline may offset the issues caused by the low quality of the textual data.

One such methodology is the Multi-Dimensional Analysis (MDA) method of register analysis (Biber 1988), which makes use of a large number of lexico-grammatical features to extract a number of dimensions of functional variation from a set of texts, such as informational versus involved production or narrative versus non-narrative concerns. In this paper, we perform an MDA of a subset of ECCO to analyze functional variation within Scottish Enlightenment writing in order to gauge the robustness of MDA when facing the sometimes extremely low quality of the ECCO OCR data.

Many of the typical features used in MDA will likely not be identified reliably in our analysis, but based on our hypothesis we expect that the features which are identified are nonetheless enough to provide meaningful results. We will evaluate the analysis and the results both quantitatively (à la Hill & Hengchen 2019) and qualitatively. Our preliminary analyses suggest that meaningful variation can be identified even in lower-quality data. Earlier research has also shown that regardless of the exact set of features, MDA studies consistently produce “compatible” register dimensions (McEnery & Hardie 2012: 115).

**References**

Biber, D. (1988). Variation across speech and writing. Cambridge University Press. <https://doi.org/10.1017/CBO9780511621024>

Eisenstein, J. (2013). What to do about bad language on the internet. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 359-369). <https://aclanthology.org/N13-1037>

Hill, M. J., & Hengchen, S. (2019). Quantifying the impact of dirty OCR on historical text analysis: Eighteenth Century Collections Online as a case study. Digital Scholarship in the humanities, 34(4), 825-843. <https://doi.org/10.1093/llc/fqz02>

McEnery, T., & Hardie, A. (2012). Corpus linguistics: Method, theory and practice. Cambridge University Press. <https://doi.org/10.1017/CBO9780511981395>

**Keywords:** multi-dimensional analysis (MDA), register analysis, historical linguistics, OCR quality, Eighteenth Century Collections Online (ECCO)

[7869] ***Corpus Linguistics and Media Guidelines – An International Case Study on Obesity in the News.***

Tara Coltman-Patel (Tara Coltman-Patel), Monika Bednarek (University of Sydney), Carly Bray (University of Sydney) and Catriona Bonfiglioli (University of Technology Sydney).

**Abstract.** It is widely recognised that weight bias in news media is a problem, with images, language, and news frames shown to contribute to negative and stereotypical representations of people with obesity (Brookes and Baker 2021).

In response, several international organisations, active in improving perceptions and treatment of people with obesity, have published media guidelines that aim to provide journalists with recommendations on how to effectively but sensitively report on weight and obesity.

We first review six guidelines published for national and global contexts, focusing specifically on recommended language use, and discuss the guidelines’ similarities and disparities. Taking a corpus linguistic approach, we examine the extent to which selected recommended language use has been adhered to in two corpora: one comprising Australian national newspaper articles

(3,618,530 words), and the other comprising UK national newspaper articles (9,805,484 words). The paper concentrates on the guidelines’ most consistently recommended advice – to use people-first-language (e.g. people with obesity, woman with obesity) and to avoid identity-first-language (e.g. obese person, obese woman). People-first and identity-first-language have a large number of linguistic variations. To identify their frequency and distribution in our corpora, we used CQPweb (Hardie 2012) to carry out advanced concordance searches using CQP syntax. We then assessed how well UK and Australian newspapers aligned with language guidelines on the whole and compared the behaviour of broadsheet and tabloid newspapers in each country. Our results demonstrate that despite the existence of media guidelines, in both the UK and Australia, and in both broadsheet and tabloid newspapers, discriminatory reporting on weight and obesity still prevails.

**References**

Brookes, G., & Baker, P. (2021). Obesity in the News: Language and Representation in the Press. Cambridge: Cambridge University Press.

Hardie, A. (2012). CQPweb — Combining Power, Flexibility and Usability in a Corpus Analysis Tool. International Journal of Corpus Linguistics, 17(3), 380-409.

**Keywords:** Corpus Lingustics, News Discourse, Obesity, Weight Stigma

[7911] ***Light verbs in L2 Czech.***

Andrea Hudousková (Charles University, Faculty of Arts, Institute of Czech Studies).

**Abstract.** Based on the data from the Czech learner corpora CzeSL and MERLIN, the paper focuses on the use of common light verbs (have, do/make, give, get, take) in non-native Czech. Furthermore, the use of other light verbs (bring, offer, keep, find, gain, lose, induce etc.) listed in Vallex 4.0, on-line Czech valency dictionary, is surveyed.

Light verbs are semantically weak verbs complemented mostly with a lexically specific nominal expression, thus giving rise to partly idiosyncratic V–N collocations that present difficulties from the point of view of L2 acquisition. Therefore, such expressions allow to examine the idiomaticity and native-likeness of the language of L2 speakers.

The observation of the use of the abovementioned V–N collocations by non-native speakers of Czech and its comparison with the data from SYN and ORAL, written and spoken corpora of contemporary Czech, lead to the following conclusions:

1) in non-native use semantically general verbs are overused, while more specific and native-like collocations are underrepresented;

2) the range of collocations at lower proficiency levels is restricted to several core expressions, which gradually improves at higher proficiency levels;

3) whereas lower proficiency speakers use collocations in order to achieve direct communicative goals, higher proficiency learners needs them to express abstract meanings and for academic writing;

4) errors in collocations prove that they are difficult to acquire and subject to cross-linguistic influence.

Although collocations in learner language have recently received a lot of attention (Dąbrowska, Nesselfhauf, Durrant, Ellis, Gilquin, Herbst and others), they have gone unnoticed in the domain of L2 Czech so far. The paper shows the ways in which the available corpora of Czech may draw attention of Czech teachers and learners to formulaic language and also points out an urgent need for collocations learning tools responding to the needs of non-native speakers.

**Keywords:** applied corpus linguistics, L2 acquisition, L2 Czech, collocation, light verb

[8025] ***Examining the discourse of net-zero pledges by global corporations: A corpus-assisted analysis.***

Matteo Fuoli (University of Birmingham) and Annika Beelitz (University of Birmingham).

**Abstract.** Despite global efforts to address climate change, greenhouse gas emissions continue to rise and the opportunity to limit global warming to 1.5C as outlined in the Paris Agreement is rapidly diminishing. In response to this pressing problem, numerous countries have made commitments to achieve carbon neutrality, with net-zero targets covering nearly 90% of global emissions (Climate Action Tracker 2022). Large corporations, which contribute significantly to global carbon emissions and play a key role in the world's ability to combat climate change through their investment decisions, are also increasingly making net-zero pledges. However, these pledges have been criticized for being too vague and lacking substance (Dickie and Jessop, 2022).

This study combines corpus and discourse analyses to evaluate how a selection of the world's largest companies use language in their net-zero pledges and justify their actions in response to increasing public pressure to address climate change. The study examines sustainability reports published by a subset of 100 companies from the Future 500 list and uses keyword, collocation, and concordance analyses.

The findings from the corpus are analyzed through the framework of Van Leeuwen's (2007) model of discursive legitimation. Preliminary results show that the net-zero frame is quickly becoming part of corporate discourse and that carbon majors are using aspirational language and journey metaphors to frame the goal of carbon neutrality in the same way they did with the concept of "sustainability" about two decades ago (e.g. Milne et al. 2006). Therefore, it appears that, while the end goal may have shifted from sustainability to carbon neutrality, corporate discourse on climate change remains primarily a form of ‘aspirational talk’ (Christensen et al. 2013).

**References**

Climate Action Tracker. (2022). CAT net-zero target evaluations. Retrieved from <https://climateactiontracker.org/global/cat-net-zero-target-evaluations/>

Christensen, Lars Thøger & Morsing, Mette & Thyssen, Ole. 2013. CSR as aspirational talk. Organization 20(3). 372–393.

Dickie, G. and Jessop, S. (2022, June 12). Report casts doubt on net-zero emissions pledges by big global companies. Retrieved from <https://www.reuters.com/business/sustainable-business/report-casts-doubt-net-zero-emissions-pledges-by-big-global-companies-2022-06-12/>

Milne, Markus J. & Kearins, Kate & Walton, Sara. 2006. Creating Adventures in Wonderland: The Journey Metaphor and Environmental Sustainability. Organization 13(6). 801–839.

Van Leeuwen, Theo. 2007. Legitimation in discourse and communication. Discourse & Communication 1(1). 91–112.

**Keywords:** climate change, net-zero, corporate discourse, corpus-assisted discourse analysis

[8055] ***Exploring the Transfer of Topic-Comment Structures in L2 English: A Comparison of Asian L1 Speakers.***

Yasuhiro Fujiwara (Meijo University), Takanori Iwao (Kobe City University of Foreign Studies), Hajime Ito (Kansai University of International Studies) and Kiyoshi Naka (Gifu University).

**Abstract.** This study aims to explore whether L1 speakers of topic-prominent languages, such as Chinese, Indonesian, Japanese, and Korean, transfer the topic-comment structure to their English as a second language (L2). Whilst English, a subject-prominent language, demands a strict relationship between a subject and predicate, as in the sentence "I am really busy today," topic-prominent languages, such as Japanese, permit a greater range of relationships between them, as demonstrated in the sentence "Kyou wa isogashii" (Today is busy) (Li & Thompson, 1976).

Previous studies have suggested that this topic-comment structure may transfer to some extent in L2 English (Sasaki, 1990; Shibata, 2006; Komiya, 2016). Our recent study of Japanese native speakers (Fujiwara, et al., 2022) with a large corpus data also found evidence of this transfer, although the frequency was fairly limited.

This study expands the analysis to include native speakers of other Asian languages, namely Chinese, Indonesian, and Korean, which are also classified as topic-prominent languages (Li & Thompson, 1976), with a learner corpus called the International Corpus Network of Asian Learners of English (ICNALE; Ishikawa, 2013). ICNALE comprises a large number of spoken and written samples from English learners in 10 Asian countries/regions, which can be compared in terms of learners' ability (CEFR A2, B1, B2) and first language (L1; Chinese, Indonesian, Japanese, and Korean). In this analysis, we utilized data from the B1 level of written essays and spoken monologues, the largest volume zone.

The results showed that the use of the topic-comment structure was very limited across all language modes (written/spoken) and native languages (Chinese, Indonesian, Japanese, and Korean). Although Chinese learners demonstrated a relatively low frequency of use, statistical analysis revealed that there was almost no difference among the L1 groups. Further qualitative analysis will examine any notable similarities and differences between the L1 groups.

**Keywords:** L1 transfer, topic-comment structure, Asian L1 speakers, learner corpus, second language acquisition

[8062] ***Parliamentary language profiles. How do the politicians speak?.***

Martina Berrocal (Friedrich Schiller University, Department of Slavic and Caucasus Studies) and Václav Cvrček (Charles University, Institute of Czech National Corpus).

**Abstract.** This presentation summarizes a study on linguistic profiling of Czech political parties based on the parliamentary speeches of their representatives. It attempts to answer research questions such as: Is there something like conservative, liberal, left-wing, or populist language? What are their main linguistic and lexical features? Numerous studies aiming to provide a description of political (often populist) leadership discourse (for example, Zienkowski & Breeze 2019, Wodak et al 2021, Ilie 2020) have appeared in the past two decades. These were, for the most part, qualitative with a focus on rhetoric and discursive strategies. We take a different (quantitative and corpus-driven) approach. By linguistic profile, we understand a configuration of the linguistic form (linguistic features) and prominent lexical items. Since the linguistic profile of a party is a multi-faceted phenomenon, we combined two methods in our approach to target different linguistic characteristics: additive multidimensional analysis (MDA) (Sardinha et al. 2019) to capture register-related characteristics (i.e. grammatical and formal features) and keyword analysis (KWA) (Scott & Tribble 2006), to reflect lexical specificities). This methodological triangulation exploits the synergic effects of MDA and KWA (Berrocal & Cvrček 2022) and allows for a more robust data basis and a more reliable interpretation. Our corpus is a random sample of 4295 transcripts from the Parlcorp corpus (Berrocal & Berrocal 2021) that includes parliamentary speeches from 1991-2021 and the Koditex corpus (Zasina et al. 2018) as a reference corpus for the KW extraction.

**References**

Berrocal, M. & Berrocal, M. (2021). ParlCorp: Corpus of Czech Parliamentary speeches. Institute of Czech National Corpus, Faculty of Arts, Charles University, Prag. <http://www.korpus.cz>

Berrocal, M. & Cvrček, V. (2022). Combining multidimensional and keyword analysis: a way to determine topic and register keywords. CADS 2022. Bertinoro. Italy.

Ilie, C. (2020). Discussion, dispute or controversy? Paradigms of conflict-driven parliamentary practices. Journal of Language Aggression and Conflict, 9(2), 237–270.

Sardinha, T. B., & Veirano Pinto, M., Mayer, C., Zuppardi, C. & Kauffmann, C. H. (2019). "Adding Registers to a Previous Multi-dimensional Analysis." In T. B. Sardinha & M. Veirano Pinto (Eds.), Multi-dimensional analysis: Research methods and current issues (pp. 165-186). London: Bloomsbury Publishing.

Scott, M., & Tribble, Ch. (2006). Textual Patterns: Key Words and Corpus Analysis in Language Education. Philadelphia: John Benjamins. <https://doi.org/10.1075/scl.22>.

Wodak, R., Culpeper, J., & Semino, E. (2021). Shameless normalisation of impoliteness: Berlusconi’s and Trump’s press conferences. Discourse & Society, 32(3), 369–393. <https://doi.org/10.1177/0957926520977217>

Zasina, A. J., Lukeš, D., Komrsková, Z., Poukarová, P. & Řehořková, A. (2018). Koditex: A corpus of diversified texts. Institute of the Czech National Corpus, Faculty of Arts, Charles University, Prague. [www.korpus.cz](http://www.korpus.cz)

Zienkowski, J. & Breeze, R. (Eds.) (2019). Imagining the Peoples of Europe. Populist discourses across the political spectrum. Amsterdam/Philadelphia: John Benjamins.

**Keywords:** multidimensional analysis, parliamentary discourse, keyword analysis, register

[8067] ***Self-identification in GenZ environmental associations: a corpus assisted analysis of first-person plural pro-forms.***

Mariasophia Falcone (Università di Modena e Reggio Emilia).

**Abstract.** The fight against climate change has recently gained particular momentum, leading to a rise in environmental protests all over the world. From the early Friday school strikes by Greta Thunberg, these protests have seen a great deal of spontaneous youth participation (particularly members of Generation Z) with a consequent rise in environmental youth organizations.

The popularity of environmental themes within this generation and their high level of environmental awareness (Seemiller & Grace, 2018), along with a peculiar use of digital tools (Parker & Igielnik, 2020), has favoured the presence of these associations on social media, particularly Instagram.

This paper aims to investigate self-identification of GenZ activists by analysing the use of first-person plural pro-forms (i.e., we, us, our, ours) in a corpus 104.090 Instagram captions by 3 environmental youth organizations. Pro-forms were analysed in their inclusive and exclusive function (Levinson, 1983) to investigate the shift in activist association with different communities. The methodological approach combined quantitative corpus-based techniques and follow-up qualitative analysis of the pro-forms within their context of usage, resulting in the analysis of a random sample of 200 collocations for each pro-form. From the analysis, 5 categories of identification were found, within a spectrum of most-exclusive, meaning the association itself, to most-inclusive, meaning humanity along with other species in the ecosystem. Results were further analysed in relation to the three bases of collective identity (group cohesion, emotional attachment, and solidarity) (Polletta & Jasper, 2001), as applied to social media activism by Davis et al. (2019). While, on the one hand, findings were in line with recent literature on GenZ's view of climate change issues (Seemiller & Grace, 2018; Clayton, 2020), a further investigation on their construction of collective identity could bridge the gap between social media and collective identity formation in these new environmentalist movements (Gerbaudo, & Treré, 2015).

**Keywords:** GenZ, Climate change, Collective identity, Social media, Pro-forms

[8079] ***A Corpus-Based Analysis of I think in Taiwanese EFL Learners and English Native Speakers: Research from the ICNALE Spoken Data.***

Ming Hung Kao (National Chengchi University).

**Abstract.** This study investigates and compares the use of *I think* in spoken monologues by Taiwanese EFL learners and English native speakers in the corpus ICNALE (The International Corpus Network of Asian Learners of English). The main research questions are: (1) How frequently do Taiwanese EFL learners and English native speakers use the phrase *I think* in spoken monologues? (2) What are the functions of *I think* used by Taiwanese EFL learners and English native speakers? The study analyzed the frequency of *I think* based on the framework of Verdaguer (2015) and Fortanet (2004) and applied AntConc 3.5.9 software to extract each occurrence of *I think* produced by Taiwanese EFL learners and English native speakers. Each *I think* was categorized based on its function in context: cogitation, opinion, or hesitation. The findings show that Taiwanese EFL learners use more *I think* in their speeches than English native speakers (279 vs. 169). This finding may be consistent with Xu and Xu’s study (2007), in which they found that EFL learners such as Chinese speakers tend to overuse I-perspective expression to express their opinions while English Native speakers prefer to use you-perspective.

The results also demonstrate that the opinion function of *I think* is most frequent in both sup-corpora. In addition, both sub-corpora display that the conjunctions (e.g., and, so) are often associated with the opinion function of *I think* to convey their perspectives. Moreover, the hesitation pauses (e.g., um, uh) and repeated words coming with *I think* show that speakers are not sure of what to say or how to say in the hesitation function. This study provides critical pedagogical implications for raising learners’ awareness of the use of *I think*.

**Keywords:** I think, ICNALE, Learner Corpus

[8190] ***Deconstructing Discourses – A network-based approach to analysing Lone Wolf Terrorism in British Newspapers.***

Hanna Schmueck (Lancaster University).

**Abstract.** The aim of this paper is to showcase interdisciplinary novel approaches to collocation network analysis via a case study on lone wolf terrorism discourses in the British press using the Lone Wolf Corpus (LWC; Malone, 2020). Aboutness, special functions of words, and semantic relations are highly relevant when examining discursively complex topics such as lone wolf terrorism, and all of these features emerge from collocational profiles (Pecina, 2010; Xiao & McEnery, 2006; Brezina, 2016; Baker, 2016; Brezina et al., 2015).

Extending this line of thought, an analysis of clusters containing heavily interrelated collocates can represent different layers of discourses contained in a corpus and thus constitute a valuable addition to the linguist’s toolbox.

The LWC contains just under 8.5 million words of British newspaper coverage on lone wolf terrorism and serves as the basis for discourse cluster generation. spaCy (Montani et al., 2022) is used for POS-tagging and lemmatising the texts, collocation scores for all sentence-level bigrams (CPN: 6a-LogLikelihood(50), sentence, C0-NC0; filtered to only contain nouns, lexical verbs, adverbs, and adjectives[1]) are then generated in Python. The graph theoretical visualisation and analysis of the remaining list of collocations is carried out via py4cytoscape (Shannon et. al, 2003). Lastly, clusters are extracted using MCODE (Bader & Hogue, 2003) and the Girvan-Newman fast greedy clustering algorithm (Girvan & Newman, 2002). Emerging clusters are visualised using edge weighted spring directed layouts (Kamada & Kawai, 1989).

The results show that topically similar clusters (here: clusters containing temporal and special information related to terrorist attacks, names of terrorists and victims, weapons, and organisations) emerge across the different subsections of the LWC. The observed clusters present a first step towards more holistic insights into discursive patterns while moving away from an exclusively intuition-based starting point for linguistic analysis (Castro & Siew, 2020; Sinclair & Coulthard, 1975; Dong & Buckingham, 2018) .This underlines the notion that the application of graph theoretical methods to language data displays both theoretical and methodological potential (Vitevitch & Goldstein, 2014).

[1] Here taken to represent the lexicalised end of the lexicogrammatical spectrum.

\_\_\_\_

**References**

Bader, G. D., & Hogue, C. W. V. (2003). An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics, 4(2). <https://doi.org/10.1186/1471-2105-4-2>

Baker, P. (2016). The shapes of collocation. International Journal of Corpus Linguistics, 21(2), 139–164. <https://doi.org/10.1075/ijcl.21.2.01bak>

Brezina, V. (2016). Collocation Networks: Exploring Associations in Discourse. In P. Baker & J. Egbert (Eds.), Routledge Advances in Corpus Linguistics: Vol. 17. Triangulating methodological approaches in corpus-linguistic research (pp. 90–107). Routledge.

Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. International Journal of Corpus Linguistics, 20(2), 139–173. <https://doi.org/10.1075/ijcl.20.2.01bre>

Castro, N., & Siew, C. S. Q. (2020). Contributions of modern network science to the cognitive sciences: Revisiting research spirals of representation and process. Proceedings. Mathematical, Physical, and Engineering Sciences, 476(2238), 1–25. <https://doi.org/10.1098/rspa.2019.0825>

Dong, J., & Buckingham, L. (2018). The collocation networks of stance phrases. Journal of English for Academic Purposes, 36, 119–131. <https://doi.org/10.1016/j.jeap.2018.10.004>

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. Proceedings of the National Academy of Sciences of the United States of America, 99(12), 7821–7826. <https://doi.org/10.1073/pnas.122653799>

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>

Malone, D. (2020). Developing a complex query to build a specialised corpus: Reducing the issue of polysemous query terms. Corpora and Discourse International Conference 2020. Advance online publication. <https://doi.org/10.13140/RG.2.2.31214.43846>

Montani, I., Honnibal, M., van Landeghem, S., & Boyd, A. (2022). spaCy: v3.4.0 [Computer software].

Pecina, P. (2010). Lexical association measures and collocation extraction. Language Resources & Evaluation, 44, 137–158. <https://doi.org/10.1007/s10579-009-9101-4>

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>

Sinclair, J., & Coulthard, M. (1975). Towards an analysis of discourse: The English used by teachers and pupils. Oxford University Press.

Vitevitch, M. S., & Goldstein, R. (2014). Keywords in the mental lexicon. Journal of Memory and Language, 73, 131–147. <https://doi.org/10.1016/j.jml.2014.03.005>

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>

**Keywords:** cluster analysis, collocations, terrorism, graph theory, lone wolf, network, British press

[8214] ***The use of internal states words by bilingual children with Down syndrome: Longitudinal development, cross-linguistic variation, and register variation.***

Jesse Egbert (Northern Arizona University), Gaëtanelle Gilquin (UCLouvain) and Tove Larsson (Northern Arizona University).

**Abstract.** Verbal communication is a major challenge for individuals with Down syndrome (DS) (Chapman et al., 1991). Children with DS tend to lag behind typically developing children (Martin et al., 2009). This lag is evident in the abilities of children with DS to produce internal states words, including lexical items used to express sensory perception (e.g. vision, hearing, smell), affect (e.g. affection, pleasure, sadness), and cognition (e.g. knowledge, thought, memory) (Beeghly & Cicchetti, 1997).

This pattern of delayed development may be compounded by the acquisition of two languages in bilingual children with DS (Cleave et al., 2012). Building on previous research, we explore longitudinal development in bilingual children with DS in two languages (English and French) and two registers (oral narratives and conversation). Our research questions are:

1. To what extent does the use of internal states words by children with DS change over time?

2. To what extent does the use of internal states words by children with DS differ across languages (English and French)?

3. To what extent does the use of internal states words by children with DS vary across registers?

Participants were English-French bilingual children with DS (n = 25) who produced elicited oral narratives and conversation with adult interlocutors at six-month intervals over approximately two years (see Cleave et al., 2012).

We adopted a corpus-driven approach to analyzing internal states words in which we generated word frequency lists for the entire corpus and identified words that were potential candidates for internal states words by examining the words in context, coding individual words, and measuring inter-rater reliability. Quantitative and qualitative comparisons were conducted across time points, between languages, and between registers. Results reveal measurable effects of each of these variables. Our findings have implications for researchers and professionals in communication disorders and caregivers of children with DS.

**Keywords:** Down syndrome, communication disorders, child language acquisition, bilingualism, longitudinal development, register variation

[8221] ***Negotiating the ‘good’ and ‘bad’ self: comparing representations of people with BPD on the r/BPD and r/BPDLovedones subreddits.***

James Balfour (University of Glasgow).

**Abstract.** ‘Seeing people with BPD as fire instead of monsters helped me immensely’, writes one user on the BPDlovedones subreddit, ‘they are dangerous and will hurt us if we get too close’. The BPDLovedones subreddit is an online forum where family, friends and other loved ones of people diagnosed with borderline personality disorder (‘BPD’) share narratives and seek advice on how to respond to people with the disorder.

The subreddit, which, at the time of writing, has over 6.81K members, presents itself as seeking to 'help users process and understand the confusion, frustration and pain that can arise from people involved with someone who has this disorder.'

The lack of public awareness about the disorder and the fact that symptoms of BPD, a personality disorder, are not necessarily immediately perceived as pathological, explains the unusual popularity and activity on this site. The DSM V describes BPD as a complex disorder characterised by “a pervasive pattern of instability of interpersonal relationships, self-image, and marked impulsivity” (APA, 2013:663). People with BPD are one of the most vulnerable groups in society with more than 60% of those diagnosed attempting suicide in their lifetimes (Soderholm et al, 2020).

This paper identifies key issues raised in the r/BPDLovedones subreddit and compares how the same issues are linguistically framed in another online forum, r/BPD, which instead almost exclusively exhibits activity from people diagnosed with the disorder. To carry out the analysis, a corpus of posts on the subreddit between 1 December and 14 December 2020 was constructed (totalling 713, 317 tokens). Using the Spoken BNC 2014 as a reference corpus, this study generated 96 keywords (ordered by Log Ratio score and statistical significance), which were then examined in more qualitative detail. Meanwhile, a comparable corpus was constructed from the r/BPD subreddit (containing 816,763 tokens), which was then used to compare lexiogrammatical patterns around the same lexical items. A key semantic category which emerged in r/BPDLovedones was “selfhood”. Somewhat surprisingly, references to selfhood were situated amidst intersection of discourses of authenticity and morality (e.g. references to good/bad and true/real/fake selves, respectively) and users made use of metaphorical framings (primarily container metaphors) to position either the ‘good’ self or ‘bad’ self as authentic. In contrast, users of the r/BPD forum typically used language to frame their disorder as an individuated grammatical entity acting upon their real and authentic self. Thus, an interesting distinction emerged between the two datasets. While a majority of users on the r/BPDLovedones subreddit linguistically equated the individual with BPD with their illness, that is the ‘bad’ self, typically, those diagnosed with the disorder, and using the r/BPD subreddit, identified with the ‘good’ self. The practical implications of this finding are then discussed.

**References**

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (5th ed.). Arlington, Virginia: American Psychiatric Publishing.

Söderholm, J., Socada, J., Rosenström, T., Ekelund, J., & Isometsä, E. (2020). Borderline Personality Disorder With Depression Confers Significant Risk of Suicidal Behavior in Mood Disorder Patients-A Comparative Study. Frontiers in Psychiatry, 11, 290.

**Keywords:** Critical Discourse Analysis, Health communication, Computer Mediated Communication, Mental Health

[8233] ***Functional language in academic lectures – validating corpus linguistic findings for language assessment.***

Haoshan Ren (Georgia State University).

**Abstract.** Academic lectures have been one of the most studied spoken academic genres, especially from a corpus linguistics perspective (e.g., Deroey, 2012; Deroey & Taverniers, 2012; Chang, 2012). However, existent literature covers a wide variety of linguistic units studied under different functional frameworks, which complicate the process of applying current findings to future research, assessment, or pedagogical uses. Besides, although corpus-informed language features start to emerge as a construct in language testing of non-native speaking instructors (e.g., Römer, 2017), no empirical evidence has shown how functional linguistic units relates to or interact with the other constructs, or together how they impact students’ or expert raters’ perception of lecture quality. This issue has been increasingly pertinent to the population of international teaching assistants (ITAs), whose academic trajectories are dependent on ITA placement tests that assess their teaching ability.

Therefore, this study first presents a systematic review of 26 previously published corpus studies on academic lectures in the past decade, revealing three main types of linguistic units as well as two types of functional frameworks employed in these analyses. This review provides a practical foundation for investigating how the use of the different types of functional units influences raters’ judgement of teaching effectiveness.

A sequential empirical study was conducted to model the relationship between language proficiency, functional language use, and overall teaching effectiveness. Using data from 195 ITA teaching demonstrations in an audio-recorded ITA test, a linear regression analysis was conducted using language proficiency scores and functional language indices as independent variables, and teaching effectiveness as the dependent variable. The results show that certain types of functional language use significantly influence expert raters' perception of teaching effectiveness. Results from this study provide empirical evidence for the connection between corpus linguistic studies and their applications in language assessment.

**Keywords:** Spoken corpora, academic lectures, functional language, corpus studies and language assessment, International teaching assistants

[8263] ***Professional nouns in a positive light: How are internationally standardized groups reflected in news media?***

Irene Elmerot (Stockholm university, dep. of Slavic and Baltic languages, Finnish, Dutch and German).

**Abstract.** Every twenty years, the International Labor Organization publishes a list, ISCO, of professional titles categorised into ten groups, based on socio-economic factors, where ISCO 1 is the highest ranking and 9 is the lowest, and 0 is military personnel. This study combines nouns from these groups and a group of public figures with all adjectives in the Subjectivity Lexicon for Czech (Veselovská and Bojar 2013) that are evaluated as positive or negative. The evaluations of the Subjectivity Lexicon have for this presentation been calculated, inspired by (Cvrček 2014), per co-occurrence and year to show the trends for different nouns in the corpus, and thus in Czech printed media over 30 years. The dataset contains 8.7 million co-occurrences, or sampling points (term from Gabrielatos et al. 2012, 153), from a period of 30 years and a variety of printed news media, extracted from a journalistic subcorpus of the Czech National Corpus.

The study, partially created to be methodologically reproducible for other languages, appears to be one of the few in linguistics that uses ISCO data. According to previous research outside linguistics (e.g. Nguyen and Do 2022; Figenschou, Eide, and Einervoll Nilsen 2021; Alhamdan et al. 2014; Musílek and Katrňák 2015), the groups with a lower socio-economic status, ISCO groups 7–9, are less frequently mentioned in the news, and when they are, it is more often in a negative light.

The research questions are

1. Which professional group(s) are a) more positively and negatively evaluated, and b) more frequently occurring in the dataset?

2. Which changes over time are visible for the different ISCO groups?

The findings demonstrate that the co-occurrences of the lowest professional ranks are not always less frequent or more negative. Instead, the reasons seem to lie rather in how much they either supervise others, or provide service to others in their profession. The presentation will also contain a discussion on limitations of the ISCO groups and Subjectivity Lexicon.

**Keywords:** news media, International Labour Organisation Classification of Occupations (ISCO), Czech language, subjectivity and sentiment, professional titles

[8310] ***Using a Representativeness Argument for Corpus Evaluation.***

Jenny Kemp (Leicester University).

**Abstract.** Most corpora are a sample taken from a wider domain or population, so corpus investigators extrapolate their findings to that wider domain. To enable such inferences, corpus builders should demonstrate that the sample adequately represents the content and language of the target domain in a manner which is transparent for corpus consumers. Yet the problem of representativeness is that it is often not fully investigated or reported in the literature - an issue which Egbert et al. (2022) go a long way towards addressing. In this paper, I offer a solution for vocabulary studies, namely the construction of a Representativeness Argument, which is an adaptation of Stephen Toulmin’s (1958/2003) argument structure. It is illustrated using my 1.98-million word corpus of texts from postgraduate International Law (IL), compiled in order to identify a discipline-specific vocabulary core (DSVC) for second language students of International Law.

The representativeness argument clearly shows how both the target domain and linguistic representativeness of the DSVC-IL corpus have been evaluated. Target domain representativeness was investigated by surveying academic programmes and reading lists, and using expert judgements. Linguistic representativeness was examined by assessing the stability of a sample of words, the 1026-flemma Single Word (SW) List. Ranked Average Reduced Frequency scores (Savický and Hlaváčová 2002) were compared for randomly-split corpus halves; and SW List coverage of texts within and outside the corpus was assessed. Expert judgements were also used to find out the extent to which lawyers saw list items as discipline-specific. On the basis of the representativeness argument presented, consumers can be confident in the validity and reliability of the DSVC-IL corpus and the SW List. The innovation of providing a representativeness argument in corpus research is important because it offers a way in which to justify the validity and reliability of research findings to a multidisciplinary audience in a transparent manner. It is particularly useful for ESP corpora and vocabulary research.

**References**

Egbert, J., Biber, D. & Gray, B. (2022) Designing and Evaluating Language Corpora: A Practical Framework for Corpus. Cambridge University Press.

Savický, P. & Hlaváčová, J. (2002) 'Measures of word commonness', Journal of Quantitative Linguistics, 9(3), 215-231.

Toulmin, S.E. (1958/2003) The Uses of Argument. Cambridge: Cambridge University Press.

**Keywords:** representativeness, corpus evaluation, vocabulary research, ESP corpora, ESAP corpora, word list research, expert judgements

[8345] ***Super as a cross-linguistic intensifier.***

Chad Howe (University of Georgia), Camila Lívio (University of Georgia) and Katherine Ireland (University of Georgia).

**Abstract.** The study of intensifiers in English has focused primarily on the interplay between high frequency variants, such as very tall and really tall, and the emergence of newer variants like so (Ito & Tagliamonte 2003, Tagliamonte & Roberts 2005, Brown & Tagliamonte 2012). Despite some attention in the literature (see Waksler 2012), more recent developments have been understudied, particularly with respect to their tendencies for cross-linguistic generalization. This paper examines the distribution of super as an intensifier, providing comparative corpus evidence from English, Portuguese, and Spanish that this exhibits a type of parallel patterning across these languages.

The data used in this analysis were extracted from three of the corpora available through SketchEngine (Kilgarriff et al. 2014): English Web 2020 (enTenTen20), Portuguese Web 2018 (ptTenTen2018), and Spanish Web 2018 (esTenTen18). The part of speech tagging in the corpus facilitated the process of extracting collocates by focusing on the structure SUPER + Adjective.

(1) Woah, this looks super interesting and I love the art as well (enTenTen20)

(2) Achei super interessante o projecto (ptTenTen18)

‘I find the project super interesting’

(3) La oferta laboral me parece super interesante (esTenTen20)

‘The work offer seems super interesting to me’

The comparison of the data from English (1), Portuguese (2), and Spanish (3) reveals a number of important patterns. First, comparison of the individual adjectives suggests a high rate of cross-linguistic similarity between specific cognates, as exemplified by interesting/interessante/interesante in (1-3). This effect, we argue, is characteristic parallel extension of this structure across languages, with specific lexical items serving as a vector for change. Moreover, the data from Portuguese and Spanish suggest a variable morphological suffixation that is not observed in the English data (e.g., superbueno ‘super good’ in Spanish; see Foltran & Nóbrega 2016). In sum, the use of large-scale corpus data in this paper provides a cross-linguistic view for the distribution of novel intensifiers.

**References**

Brown, L., & Tagliamonte, S. A. 2012. A Really Interesting Story: The Influence of Narrative in Linguistic Change. University of Pennsylvania Working Papers in Linguistics. 18(2): 1–10.

Foltran, M. J., & Nóbrega, V. A. 2016. Intensifier adjectives in Brazilian Portuguese: Properties, distribution, and morphological reflexes/Adjetivos intensificadores no Português Brasileiro: Propriedades, distribuição e reflexos morfológicos. Alfa: Revista de Linguística, Vol, 2, 319–340.

Ito, R., & Tagliamonte, S. A. 2003. Well weird, right dodgy, very strange, really cool: Layering and recycling in English intensifiers. Language in Society. 32:257–279.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. Lexicography. 1(1): 7–36.

Tagliamonte, S., & Roberts, C. 2005. So weird; so cool; so innovation: The use of intensifiers in the television series Friends. American Speech. 80: 280–300.

Waksler, R. 2012. Super, uber, so, and totally: Over-the-top intensification to mark subjectivity in colloquial discourse. In N. Baumgarten, I. Du Bois, & J. House (Eds.), Subjectivity in Language and Discourse, 17–31. Netherlands: Brill.

**Keywords:** variation, intensifiers, English, Portuguese, Spanish

[8349] ***Acquisition of Arabic Agreement System by Learners of Arabic as a Second Language.***

Manar Almanea (Al Imam Mohammad Ibn Saud Islamic University).

**Abstract.** This study investigates the stages of acquisition of the grammatical agreement system in Arabic by learners of Arabic as a second language. One of the tools of the study is the construction of a specialized learner corpus at varying levels of proficiency.

The general purpose is to decide whether the agreement system in Arabic, with its various subcomponents, is acquired according to a specific order. It attempts to establish the order of acquisition, if any, across the complex of agreement relations in Arabic, which are noun-adjective (N-A), subject predicate (S-P), subject-verb (S-V) and verb-subject (V-S). A descriptive, cross-sectional, developmental approach is adopted. Data was collected from 278 learners of Arabic speaking 76 L1s. They were enrolled at three academic levels at an Arabic Language Institute. Four data collection instruments have been employed. They are a completion task, a grammaticality judgment task, a picture-description task and a spontaneous writing task. Data obtained from the spontaneous writing task was digitized in the form of a specialized Arabic learner corpus. The corpus was tagged for all occurrences of agreement structures. A software concordancing program compatible with Arabic (aConCorde) was used to search through the corpus and to extract results. The data analysis was based on the obligatory occasion analysis method.

The results revealed the existence of a general order of acquisition of the four agreement relations. Learners acquired the S-V agreement relation first, followed by S-P agreement, N-A agreement, and finally V-S agreement. In addition, among all the four agreement relations, the masculine singular form was acquired first, followed by the feminine singular. The feminine plural was the most difficult form. With regard to number, singular is acquired first, followed by plural and finally by dual. As for person, 3rd person is acquired first followed by 1st person and last by 2nd person.

**Keywords:** Learner corpus, annotation, Arabic Agreement System, Subject-Verb agreement, Noun-Adjective agreement, Subject-Predicate agreement

[8354] ***Popular and Scientific Discourse on Antidepressants in Britain and China: The case of Venlafaxine.***

Fang Wang (Centre for Translation Studies, University Of Surrey), Sabine Braun (Centre for Translation Studies, University of Surrey) and Robert Geyer (Department of Politics, Philosophy and Religion, Lancaster University).

**Abstract.** Popular and professional media play a significant role in shaping the social construction of prescription medication. However, different national contexts create significant variation in these constructions – even when comparing a similar medication. To explore this, our research investigates the social construction of a prescribed antidepressant - Venlafaxine in British and Chinese newspapers and scientific journals and examines how the understanding of medication is significantly determined by how it is represented. Our work will be based on:

• British news corpus (669,340 words - articles in UK national newspapers covered by Nexis) and medical corpus (443,297 words - articles in UK medical journals covered by the Medline).

• Chinese news corpus (118,400 words - articles in key news websites) and medical corpus (1,008,349 words – articles in medical journals extracted from China National Knowledge Infrastructure).

Frequency, collocation, keyword and paraphrase analyses will be used to identify pattens of meanings. Initial findings suggest that in British news context, side-effects such as heart disease and the risks of combined use of Venlafaxine and anti-psychotic medication are emphasized. Whereas in the Chinese case, the side effect of tremors is emphasised while its sleeping benefits are promoted, which may contribute to reducing the social stigma of taking antidepressants in Chinese society. Interestingly, some of these differences are echoed in scientific discourse: in British medical journals, the risk of heart disease and the mechanism/assessment of drug interaction are more frequently portrayed than in Chinese journals. Both British and Chinese journals compare the efficacy of Venlafaxine with other types of antidepressants, but the side-effects of Venlafaxine were less discussed in Chinese journals. By contrast, the British journals discuss risk profiles of Venlafaxine in a more detailed way, including cardiovascular toxicity, arrhythmia, memory disorder, hyponatremia, etc.

Acknowledgement: This research is funded by British Academy/Leverhulme Small Research Grants (reference: SG2122\210988).

**Keywords:** Corpus Linguistics, Discourse Analysis, Antidepressants, Venlafaxine

[8355] ***“The invisible disability”: Media representations and attitudes towards dyslexia and people with dyslexia in the UK and Italy.***

Gianmarco Vignozzi (Università di Pisa) and Gloria Cappelli (Università di Pisa).

**Abstract.** The paper reports on a study of media representations of dyslexia and people with dyslexia in the UK and Italy. Attitudes towards and beliefs about this specific learning difficulty (SLD) have been investigated from several perspectives, including teachers’ (Hornstra et al. 2010), lay people (Furnham 213), and people with dyslexia themselves (Gibby-Leversuch 2021). Previous corpus-based studies dealing with the representation and coverage of commonly stigmatized conditions such as schizophrenia (Balfour 2019), obesity (Baker 2019), diabetes (Bednarek, Carr 2019) have observed that the media tend to polarize representations, often fueling misconceptions and stigmatisations. Similarly, Simblett (2021) has discussed the media portrayal of the UK

educational system as a site of failure for dyslexic people. Building on previous research, our investigation has a twofold aim. Firstly, we critically analysed recurrent lexico-grammatical patterns used by journalists to frame dyslexia and people with dyslexia in the UK and Italy. Such patterns were retrieved through corpus-assisted discourse analysis tools (cf. Partington et al. 2013) from two self-compiled comparable corpora of British and Italian newspaper articles and newscast transcripts spanning over the last 20 years and comprising almost 300,000 tokens each, i.e., dyslexia\_EN and dyslexia\_IT. Secondly, we tried to verify whether and how these representations are reflected in people’s attitudes towards this issue. To this aim, data from the two corpora were compared to data obtained through an online survey. The questionnaire was administered to English and Italian native speakers, who were asked to answer a series of open-ended and closed questions focusing on their beliefs and attitudes towards dyslexia.

**References**

Balfour, J. 2019. “‘The mythological marauding violent schizophrenic’: Using the word sketch tool to examine representations of schizophrenic people as violent in the British press”. Journal of Corpora and Discourse Studies, 2, pp. 40- 64. DOI:

Baker, P. 2019. “Analysing Representations of Obesity in the Daily Mail via Corpus and Down-Sampling Methods”, in J. Egberr and P. Baker (eds) Using Corpus Methods to Triangulate Linguistic Analysis, London/New York: Routledge.

Bednarek, M., Carr, G. 2019. “Diabetes coverage in Australian newspapers (2013-2017): A computer-based linguistic analysis”. Health Promotion Journal of Australia, 31/3, pp. 497-503. DOI: <https://doi.org/10.1002/hpja.295>

Furnham, A. 2013. “Lay knowledge of dyslexia”. Psychology, 4(12).

Gibby-Leversuch, R., Hartwell, B. K., and Wright, S. 2021. “Dyslexia, literacy difficulties and the self-perceptions of children and young people: A systematic review”. Current Psychology, 40(11), pp. 5595-5612.

Hornstra, L., Denessen, E., Bakker, J., Van Den Bergh, L., and Voeten, M. 2010. “Teacher attitudes toward dyslexia: Effects on teacher expectations and the academic achievement of students with dyslexia”. Journal of learning disabilities, 43(6), pp. 515-529.

Partington, A., Duguid, A., Taylor, C. 2013. Patterns and Meanings in Discourse: Theory and Practice in Corpus-Assisted Discourse Studies (CADS). Amsterdam/Philadelphia: John Benjamins.

Simblett, C. 2021. The social construction of dyslexia in the UK media: school as a site of failure (Doctoral dissertation, Durham University).

**Keywords:** corpus assisted discourse study, disability representation, cross-cultural study, perception study, dyslexia

[8380] ***A corpus analysis of creativity in care home interactions.***

Almut Koester (WU).

**Abstract.** The professional context of a care home is not an obvious choice to explore language and creativity. As research has shown, interactions between carers and residents in elder care institutions are highly routinized (Backhaus, 2017) due to the constraints of the work routine: exchanges are typically brief and characterized by short turns and language accompanying actions, such as administering medication, or feeding and dressing residents.

But it is precisely because of these constraints that it is interesting to explore the extent to which creative language and discourse does (or does not) occur in this workplace setting. The paper presents findings from a corpus of approximately 53,000 words (compiled from over 70 hours of audio-recordings) of naturally-occurring interactions in an elder care home in England.

Exploring creativity through the lens of corpus linguistics is also not an obvious choice, as creative language and discourse choices tend to be infrequent, and therefore not easily identifiable through quantitative corpus tools. Two methods were used to identify creative uses of language and discourse using corpus tools. First, types of linguistic items that have been identified in previous research as (potentially) creative, such as metaphors, idioms and hyperbole (Carter 2016), were investigated. Second, after identifying the frequent and key items in the corpus, infrequent or untypical items were also investigated with the aim of identifying any creative uses. As creative language is by definition unusual and departs from the norm (Hoey, 2005: Hanks, 2013, Carter 2016), focusing on infrequent rather than frequent items seemed a promising approach to explore creativity through corpus methods.

The analysis showed that creative language use in this workplace setting was indeed rare, but could be found in relationally-oriented talk which departed from the care routine and where alignment between the discourse participants was established.

**References**

Backhaus, P. (2017). Care Communication: Making a Home in a Japanese Eldercare Facility. Abingdon, Oxon: Routledge.

Carter, R. (2016). Language and Creativity. London: Routledge.

Hanks, Patrick (2013). Lexical Analysis. MIT Press: Cambridge, Massachusetts.

Hoey, Michael (2005). Lexical Priming. London: Routledge.

**Keywords:** elder care, creativity, workplace discourse

[8604] ***A Corpus-Assisted Critical Discourse Analysis of Antidepressants in British and Chinese Newspapers and Medical Journals: The case of Escitalopram.***

Fang Wang (University Of Surrey).

**Abstract.** Depression is one of the most widespread mental health issues around the world, and the consumption of antidepressant drugs have risen steadily in recent decades in Britain and China. What British and Chinese read about antidepressants in popular media and scientific discourse will contribute/constrain their ability to evaluate the risks and benefits of antidepressants. This research investigates the social constructions of escitalopram, a commonly prescribed antidepressant in British and Chinese newspapers and medical journals. Four corpora were built for this research project: British news corpus (265,644 words) includes articles in UK national newspapers covered by Nexis and medical corpus (545,969 words) comprises articles in UK medical journals covered by the Medline. Chinese news corpus (13,371 words) includes articles in key news websites and medical corpus (343,266 words) comprises articles in medical journals extracted from China National Knowledge Infrastructure). Frequency, collocation, keyword and paraphrase analyses were used to identify pattens of meanings. Initial findings suggest that in both British and Chinese news contexts, comparisons between the effects of escitalopram and psilocybin (psychedelic mushroom) on depression are overwhelmingly talked about and psilocybin is represented as superior to normal types of Selective Serotonin Reuptake Inhibitors (SSRIs). Furthermore, both British and Chinese news construct the most important side-effect of escitalopram as sex dysfunction. Other side-effects that Chinese news represent frequently include deteriorating liver function and nausea while British news discuss more about withdrawal symptoms. In British medical journals, explanations to mechanism of actions of escitalopram prevail, which contributes to prove the effectiveness of this medication while Chinese medical journals predominantly construct escitalopram as a helpful antidepressant to cope with old people’s depression.

Acknowledgement: This research is funded by British Academy/Leverhulme Small Research Grants (reference: SG2122\210988).

**Keywords:** Corpus Linguistics, Discourse Analysis, Antidepressants, Escitalopram

[8643] ***Distinguishing False News Through Lexico-Grammatical Features: A Multi-dimensional Analysis.***

Bashayer Baissa (University of Birmingham), Matteo Fuoli (University of Birmingham) and Jack Grieve (University of Birmingham).

**Abstract.**

Can false news articles circulating online be distinguished from other news registers through their lexico-grammatical features? Previous research has attempted to answer this question by comparing online false news articles to mainstream news articles, such as broadsheet articles (e.g., Horne & Adalı, 2017; Rashkin et al., 2017). However, such research does not control for register variation, which may act as a confounding variable. That is, differences between online false news articles and mainstream news articles may be due to variation across news registers rather than to the content’s veracity. This study investigates this question through a multi-dimensional analysis (MDA) (Biber, 1988) of false news and four news registers: broadsheets, tabloids, web-based publications, and news blogs. MDA first identifies dimensions of linguistic variations (i.e., systematic co-occurring patterns of lexico-grammatical features that together serve a communicative function) in registers under investigation, and then these registers are compared along the dimensions. The false news corpus comprised 101 online verifiably false news articles (94983 words) about climate change, vaccination, and COVID-19, and each comparative corpus (total 273,705 words) included 101 articles about the three topics under consideration.

The analysis revealed three dimensions: Dimension 1 represents the opposition between involved and informational reporting; Dimension 2 represents the opposition between expository and narrative reporting; Dimension 3 presents the opposition between overtly advocating and non-overtly advocating reporting. The study found that false news was similar to traditional news registers, namely broadsheets, tabloids, and web-based publications in that it uses both involved and informational reporting. Moreover, false news was similar to broadsheets in using elements of both expository and narrative reporting. However, false news was distinctive in that it did not overtly advocate for a certain point of view. Limitations and recommendations for future research are discussed.

**References**

Biber, D. (1988). Variation across Speech and Writing (1st ed.). Cambridge University

Press. <https://doi.org/10.1017/CBO9780511621024>

Horne, B., & Adalı, S. (2017). This Just In: Fake News Packs a Lot in Title, Uses

Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News.

9.

Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., & Choi, Y. (2017). Truth of Varying

Shades: Analyzing Language in Fake News and Political Fact-Checking. Proceedings

of the 2017 Conference on Empirical Methods in Natural Language Processing, 2931–

2937. <https://doi.org/10.18653/v1/D17-1317>

**Keywords:** false news, register variation, lexico-grammatical features, multi-dimensional analysis

[8648] ***Code-switching in Tunisian Arabic: a multifactorial Random Forest analysis.***

Chadi Ben Youssef (UCSB) and Stefan Th. Gries (University of California, Santa Barbara).

**Abstract.** Speakers must constantly make choices between a number of competing strategies and functional units to form their utterances (Du Bois, 2014). This is true of monolingual settings, but even more so in bilingual speech communities and diglossic societies (Ferguson, 1959), where two or more repertoires constantly compete–often leading to code-switching (CS). What motivates CS has been one of the most researched phenomena in language contact since the influential publication of Poplack (1980). However, studying CS using multifactorial corpus linguistics techniques is less common and tends to either use written data (Gambäck & Das, 2016), immigrant speech communities (Carter et al., 2010), or conversations of a limited number of speakers in intimate contexts (Myslin & Levy, 2015). Furthermore, such quantitative analyses focused mainly on high resource languages.

The present paper investigates CS between a low resource language, Tunisian Arabic, and French, in the TuniCo corpus (Dallaji et al., 2015). For comparability, we retained only naturally occurring narrative interviews in the corpus, resulting in 11 files containing 56,310 words produced by 13 speakers of different socioeconomic backgrounds, native of Tunis, and aged 35 and younger. We annotated each conversation for structural, psycholinguistic, and sociolinguistic factors including the part-of-speech and length of words, the language momentum and the normalized position of words within an utterance, as well as the speaker and the conversation. We model CS in these data with a Random Forest, which shows that CS is affected by several interacting considerations: (i) to maintain the code-integrity at the phrase and discourse levels, speakers tend to switch dependent parts-of-speech when the latter’s head is switched; (ii) NPs are a prime location for CS; and (iii) when speakers code-switch, they are attuned to the cognitive load they impose on themselves and/or on listeners.

**References**

Carter, D., Davies, P., Couto, M. D. C. P., & Deuchar, M. (2010). A corpus-based analysis of codeswitching patterns in bilingual communities. Revista Española de Lingüística.

Dallaji, I., Gabsi, I., & Procházka, S. (2015). TUNICO Linguistic Dynamics in The Greater Tunis Area. <https://tunico.acdh.oeaw.ac.at/index.html> [Last accessed 12/03/19]

Du Bois, J. W. (2014). Towards a dialogic syntax. Cognitive Linguistics 25(3). 359–410.

Ferguson, C. A. (1959). Diglossia. Word 15(2). 325–340.

Gambäck, B., & Das, A. (2016). Comparing the Level of Code-Switching in Corpora. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16), 1850–1855.

Myslín, M., & Levy, R. (2015). Code-switching and predictability of meaning in discourse. Language 91(4), 871–905.

Poplack, S. (1980). Sometimes I’ll start a sentence in Spanish Y TERMINO EN ESPAÑOL: Toward a typology of code-switching. Linguistics 18(7-8). 581-618.

**Keywords:** Code-Switching, Tunisian Arabic, French, Random Forest, Code Integrity, Code Momentum

[8660] ***Perceptions of Using Corpora to Enhance L2 Interactional Competence: A Study of Chinese L1 Speakers.***

Sarah Wafield (NYU Shanghai).

**Abstract.** This study aims to investigate the effectiveness of using spoken corpora to improve second language (L2) learners' interactional competence. The present study will examine the effects of corpus-based tasks on L1 speakers of Chinese enrolled in an advanced EAP course at an EMI Sino-Western university. As part of the course, learners will receive training on how to use the Trinity Lancaster Corpus (TLC) and create a specialized data set of their own group discussions. They will also engage in noticing tasks comparing data from the TLC to their class data set. Participants will provide feedback on the use of corpora to improve their turn-taking skills during group discussions in Academic English. Data will be collected through perception surveys assessing the perceived effectiveness of corpus-based methods, their impact on L2 interactional competence, and the

likelihood of using corpora in future academic endeavors. The results of this study will shed light on the usefulness of corpora for improving L2 interactional competence and may inform future research and pedagogy in the field.

**Keywords:** interactional competence, corpus-based teaching, L2 spoken corpora, English for academic purposes (EAP), Trinity Lancaster Corpus (TLC), turn-taking skills

[8695] ***The effect of English extramural activities on L2 students’ lexical diversity and grammatical complexity.***

Henrik Kaatari (University of Gävle), Tove Larsson (Northern Arizona University), Ying Wang (Karlstad University), Seda Acikara Eickhoff (Northern Arizona University) and Pia Sundqvist (University of Oslo).

**Abstract.** Frequent engagement in extramural English (EE) activities (i.e., self-initiated out-of-school English-language activities) are found to positively influence L2 high school students’ vocabulary size (Sundqvist 2009). However, as studies of this kind have mainly looked at students’ receptive knowledge, our understanding of the relationship between EE activities and students’ production remains somewhat rudimentary. Specifically, the possible impact on grammar and lexis has received very limited focus. The present study examines the effect of EE activities on lexical diversity (operationalized as moving average type-token ratio) and NP complexity (attributive adjectives and post-modifying prepositional phrases) in L2 high school student writing. The following research questions are investigated:

• What effect (if any) does EE activities have on lexical diversity and/or NP complexity?

• Are there differences between purely receptive EE activities (e.g., reading) and other types of EE activities (e.g., gaming) in terms of the effect of lexical diversity and NP complexity, and, if so, what are the differences?

Based on what we know from receptive studies (e.g., Prophète et al., 2022), we hypothesized that EE activities will have a positive effect on both lexical diversity and grammatical complexity. A subsample of the Swedish Learner English Corpus (SLEC; Kaatari et al., forthcoming) is used (grades 10–11; n=200). SLEC contains detailed information about how many hours per week students engage in five different EE activities: reading, viewing, speaking, using social media, and gaming.

We applied measured variable path analysis from the Structural Equation Modeling framework (see Larsson et al., 2021). The best-fitting model (χ2: 16.8, df: 15, CFI: 96.6, RMSEA: 0.023[0.00–0.067], SRMR: 0.039) confirmed our hypothesis that participation in EE activities has a (mostly) positive effect on both lexical diversity and NP complexity. Furthermore, the purely receptive activities (in particular reading) had an effect on lexical diversity, while the other EE activities did not.

**Keywords:** Learner corpus research, Extramural English, Lexical complexity, Grammatical complexity

[8730] ***A corpus of Victorian anti-vaccination discourse.***

Alice Deignan (University of Leeds), Tara Coltman-Patel (Lancaster University), William Dance (Lancaster University), Zsófia Demjén (University College London), Derek Gatherer (Lancaster University), Claire Hardaker (Lancaster University), Chris Sanderson (Lancaster University) and Elena Semino (Lancaster University).

**Abstract.** The Covid-19 pandemic has drawn renewed attention to the substantial minority of the population who are vaccine hesitant, an issue that also surfaced in recent decades with concerns around the HPV and MMR vaccines. However, vaccine hesitancy is not a new phenomenon, and in the UK it is traced back to the first vaccination campaign, for smallpox, in the 19th century (Durbach, 2005).

In the 21st century, the non-specialist public access anti-vaccination information and disinformation largely through online social media fora. The extant Victorian vaccination literature is from a range of formal and informal genres, of which the general public had ready access to two: the pamphlet and the informal journal, both produced quickly and cheaply, distributed freely, and handed from person to person. They could be written by anyone with the necessary literacy skills, anonymously.

In the same way, today's social media posts can be produced quickly, anonymously and with little or no cost, can be forwarded from person to person, and the content can be produced by anyone with the necessary skills, without any requirement for expertise. We therefore consider these to have closer parallels to today's social media than texts such as medical correspondence and Hansard. They are also of interest as they tend to be under-represented in historical corpora.

Our presentation describes the development of a 3.6 million word corpus of pamphlets and related texts, the Victorian Anti-Vaccination Discourse Corpus (VicVaDis). We outline the process of identifying and sourcing texts to build the corpus. We used Sketch Engine tools, principally Wordlist, Word Sketch and Concordance, to identify the arguments against vaccination that were presented to the general public. We show that there are some parallels with the anti-vaccination arguments used now, as well as additional dimensions such as religion and class struggle.

**Keywords:** historical corpus, vaccination discourse, public discourse

[8797] ***Talking to an imagined interlocutor: Interactional and interpersonal features of discourse in computer-mediated semi-direct speaking assessment.***

Dana Gablasova (Lancaster University), Luke Harding (Lancaster University), Vaclav Brezina (Lancaster University) and Jamie Dunlea (British Council).

**Abstract.** Assessing speaking is complex both in terms of defining the construct of oral proficiency and capturing that construct through suitable assessment tasks. Two major test formats have been used to assess speaking to date: direct tests, which involve face-to-face interaction, and semi-direct tests, in which test-takers react to prompts presented via different methods. Semi-direct speaking tests now feature in many computer-based English language proficiency exams (e.g. TOEFL iBT, Aptis, Pearson Test of English Academic). Due to their monologic nature, these tests have been criticised for their limited ability to capture key elements of interactional competence. However, only a few studies so far investigated the use of these pragmatic features in computer-mediated semi-direct tests (e.g. Quaid & Barrett, 2020).

To address this question, this study explores the occurrence of interaction/interpersonal features in the speaking component of a semi-direct computer-administered test, the Aptis General test (British Council, 2020), an international examination of English developed and administered by the British Council. The study draws on data from the spoken part of the BC-Lancaster Aptis corpus, which contains 650,000 words from 841 L2 English speakers at three proficiency levels (CEFR B1, B2 and C). The study focused on three sets of interactional/interpersonal markers: (a) second person personal pronouns, (b) pragmatic markers used to guide listeners’ interpretation (e.g. actually, basically), and (c) pragmatic markers used to manage information flow between interlocutors (oh, okay, right).

All three sets of interactional/interpersonal features were very frequent in the corpus and increased significantly with proficiency level. Second person personal pronouns were particularly salient, with the two categories of pragmatic markers used predominately by more proficient L2 users.

Implications are drawn for the understanding of the nature of spoken production elicited in computer-mediated semi-direct tests, where some speakers orientate themselves towards an imagined interlocutor, even in their monologic production.

**Keywords:** interactional competence, language assessment, L2 spoken production, pragmatic ability

[9004] ***Corpus approaches to transparency in corporate discourse - An analysis of texts from websites of companies in the car transportation sector.***

Federico Zaupa (University of Modena and Reggio Emilia).

**Abstract.** Within the last twenty years, significant studies applying the methods of corpus linguistics (CL) to study corporate discourse, especially Corporate Social Responsibility (CSR) communication, have mainly explored the strategic and rhetorical nature of discourse markers (see, e.g., Hyland 1998; Garzone 2005; Camiciottoli 2010), the discursive construction of corporate identity in annual reports (see, e.g., Malavasi 2010; Breeze 2013) and crucial aspects, such as trust (e.g., Fuoli 2018) and sustainability (see, e.g., Lischinsky 2010; Catenaccio 2022), as well as the extent to which the

integration of CL methods represent a valuable tool for cross-cultural genre analysis of CSR reports (see, e.g., Yu & Bondi 2017; Bondi 2022). However, one of the notions still poorly examined in linguistic research is that transparency, apart from some genre-based studies of communication and disclosure policies (see, e.g., Koskela 2018; Koskela & Camiciottoli 2020).

Drawing on this background, this paper analyses a corpus of texts (approximately 1 million of tokens) – news releases, about-us pages, and CSR sections and reports published between 2020 and 2022 – from websites of companies working in the car transportation sector (Uber, BlaBlaCar, Lyft, Europcar Mobility Group, Avis Budget Group, Zipcar).

Corpus-driven and based approaches are adopted to explore the discursive construction of transparency in this specific sector, as well as identify the main frames through which the concept of transparency is constructed, aiming at both contributing to the definition of this floating notion and providing suggestions for further linguistic research on corporate transparency across different sectors.

Quantitative and qualitative results – in particular, the lexical and phraseological patterns around the signifier transparen\* and related words identified by Koskela and Crawford Camiciottoli (2020) – suggest that transparency is largely used in general and vague terms to frame it as one of the many values the companies are committed to and that needs improvements. Findings also show that transparency is mostly associated with accountability, prices, and partly with companies’ outcomes and information disclosures, rather than sustainability and human rights issues. Based on these considerations, it further emerges that companies under scrutiny orient themselves towards organisational transparency – i.e. their reputation as transparent actors – by vague and general means of communicative transparency (Rawlins 2009).

**Keywords:** Corpus Linguistics, Discourse analysis, CSR Communication, Transparency, Car Transportation

[9025] ***Overhauling keyness analyses with three dimensions & word2vec: Contrasting three Asian varieties of English.***

Stefan Th. Gries (UC Santa Barbara & JLU Giessen).

**Abstract.** One central corpus-linguistic methods is keyness analysis, the identification and interpretation of words of a target corpus (T) that, compared to their occurrence in a reference corpus (R), are key/characteristic for T. Statistically, this is most often done with log-likelihood ratio tests comparing the token frequencies of each word type in T and R. In terms of application practice, T and R usually differ regarding their topic areas (e.g. to identify expressions relevant for topic areas in language teaching contexts) or registers/genres (e.g. to identify expressions typical of certain communicative contexts/situations).

However, ever since, e.g., Leech & Fallon (1992), studies have also used keywords analysis to compare corpora representative for different cultures; Leech & Fallon (1992) themselves compared BrE & AmE, more recent work has targeted varieties of English (e.g. Mukherjee & Bernaisch 2015 or Collins 2021).

We outline a set of suggestions for better keyness analyses and exemplify it on the basis of a cultural keyness analysis of newspaper corpora for IndE, PakE, and SriLE. We distinguish two keywords scenarios: (i) the prototype, which identifies keywords in a bottom-up fashion and (ii) one where a researcher is interested in predefined set of words and their differences across corpora. As for (i), we discuss an extension of Gries (2021) new way to compute keywords, which considers frequency but also frequency- and a dispersion-based components of keyness; we will discuss the results of comparing each variety to the other two in the SAVE 2020 corpus.

As for (ii), we exemplify how one can use distributional-semantics kinds of approaches to either disambiguate keywords yielded by the first approach or to determine how words used in each variety are associated with different kinds of ‘collocates’ (as returned by a collocational interpretation of cosine similarity) and what that indicates. Replicating Mukherjee & Bernaisch (2015), we show how the IndE, PakE, and SriE newspaper corpora differ substantially in how they discuss terror:

IndE has an overrepresentation of military terms such as artillery, combating, commando, drone, enemy, grenade, squads, …; the IndE news coverage of terror is concerned a lot with how terror is performed and what (military) reactions it prompts;

PakE has an overrepresentation of terms that involve misleading communication such as indoctrination, intimidation, misinformation, propaganda, provocation, ruse, spout, and stereotyping;

SriE has an overrepresentation of religious terms (e.g., churches, cult, fanatic, muslim, , shiite, sunni, taslim, religiosity, salafi, salafist, and extremists); the SriE news coverage considers (extreme) religiousness as relevant to terror.

**References**

Collins, P. 2021. Cultural keywords in World Englishes: A GloWbE-based study. ICAME Journal, 45, 5–35.

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.

Leech, G. & Fallon, R. 1992. Computer corpora – What do they tell us about culture? ICAME Journal, 16, 29–50.

Mukherjee, J. & Bernaisch, T. 2015. Cultural keywords in context: A pilot study of linguistic acculturation in South Asian Englishes. In Collins, P. (ed.), Grammatical Change in English World-Wide, 411–35. Amsterdam: Benjamins.

**Keywords:** keywords, frequency, dispersion, collocation, word2vec

[9089] ***Effects of Data-Driven Learning on Memorizing Verb-Noun Collocations.***

Yoshiho Satake (Aoyama Gakuin University).

**Abstract.** Appropriate collocation use is necessary for native-like language use. However, appropriate use of verb-noun collocations is difficult for learners of English as a foreign language. Data-driven learning (DDL), in which learners refer to a corpus for language learning, has been reported to be effective for collocation learning. Therefore, this study explores the effects of DDL on memorizing verb-noun collocations and enriching their associations. Additionally, this study compared the impact of learning appropriate verbs by searching for nouns or verbs in a corpus when learning verb-noun collocations.

A pre-post design evaluated the effectiveness of DDL on learning verb-noun collocations. The participants were 24 Japanese undergraduates at a private university in Tokyo, Japan, who were intermediate English learners at the B1 level in the Common European Framework of Reference for languages (CEFR). For eight weeks, the participants spent 15 minutes on nine nouns and seven verbs, learning two verb-noun collocations each week. They searched the corpus for collocations of the target words, read the concordance lines of the example sentences, and wrote down what they noticed about verbs or nouns that are often used in combination with the target words and the examples they consulted.

Pre- and post-treatment tests were administered, and the participants were asked to write appropriate collocations and associations for the words they learned. The results show that corpus consultation was significantly effective in memorizing various verb-noun collocations, with a close-to-large effect size.

However, no significant effect was detected in enriching the associations of target words. Participants wrote more appropriate collocations by writing verbs when searching for target nouns than when searching for target verbs. Instead, they wrote many light verbs, though appropriate collocation with polysemous ones might have been easier. This study suggests that DDL is a good option for learning verb-noun collocations.

**Keywords:** data-driven learning (DDL), verb-noun collocations, Teaching English as a Foreign Language (TEFL)

[9137] ***Application of collostructional analysis to describe linguistic remix.***

Emese K. Molnár (Eötvös Loránd University (ELTE), National Laboratory for Digital Heritage).

**Abstract.** There is a consensus among linguists that creativity is essential in language, therefore it is crucial to find methodological tools that help to measure the creative potential of language. The present paper provides a possible solution by adapting collostructional analysis (Stefanowitsch–Gries 2003) to Remix Studies perspective (Navas–Gallagher–burrough 2021). The proposed method is based on the comparison of typical constructions of a particular text with instances of the same construction from a general corpus. The procedure involves collolexeme analysis to determine which lemmas are most attracted to the construction in the general corpus, then the resulting lemmas are compared with previously extracted instances of the investigated text on the basis of their semantic features. To test the method, a specialized corpus of 1500 tokens was compiled from a creative language practice, in which members of the Hungarian hardcore techno community post sentences with embedded song titles on Facebook by using YouTube links of hardcore songs.

I interpret the practice as a linguistic remix (Knobel–Lankshear 2008, Lessig 2008, Williams 2017), which is considered the combination and manipulation of well-rehearsed and entrenched language units to create new ones. The most typical remix constructions of the discourse were identified as a personal pronoun + copula + noun/adjective pattern based on the POS tagging of the corpus. This copula construction was extracted from The English Web Corpus (enTenTen) and the collolexeme analysis resulted in the set of nouns and adjectives that are significantly associated with the construction in general.

By comparing the nouns and adjectives of the copula construction from both corpora based on their semantic features, it becomes possible to determine how typical or novel the remixes of the hardcore community, taking into account a general linguistic pattern. This method can be applied to a corpus-based and comparative style analysis.

**References**

Knobel, Michele – Lankshear, Colin. 2008. Remix. The art and craft of endless hybridization. Journal of Adolescent & Adult Literacy 52: 22–33.

Lessig, Lawrence 2008. Remix. Making art and commerce thrive in the hybrid economy. United States: Penguin Press.

Navas, Eduardo – Gallagher, Owen – burrough, xtine 2021. The Routledge handbook of Remix Studies and Digital Humanities. New York: Routledge.

Stefanowitsch, Anatol – Gries, Stefan Th. 2003. Collostructions. Investigating the interaction of words and constructions. International Journal of Corpus Linguistics 8: 209–243.

Williams, Quentin 2017. Remix Multilingualism: Hip Hop, Ethnography and Performing Marginalized Voices. London: Bloomsbury.

**Keywords:** collostructional analysis, linguistic remix, Remix Studies, (linguistic) creativity

[9143] ***Developing arguments and expressing subjectivity through modal devices in a local academic learner corpus: A small-scale study of key emergent language from a Moodle forum.***

Emanuela Tenca (University of Verona).

**Abstract.** This presentation focuses on the results of a small-scale study on learner output obtained by means of a Moodle forum during an English course at a B2 level, targeting students of Primary Teacher Education at the University of Modena and Reggio Emilia (Italy) in spring 2020. The activity promoted learners’ expression about topics relevant to their disciplinary and professional domain, with a view to boosting their engagement in the learning process.

The texts submitted by the students were collected after the end of the course to compile a local academic learner corpus totalling 27,430 tokens. The data gave the opportunity to explore patterns in learner discourse representing the use of modal devices (MDs) to develop arguments and express subjectivity.

In particular, this presentation is informed by two questions:

1. how are MDs distributed in the small-scale corpus?

2. what type of meanings do the MDs in the corpus express?

The corpus was investigated using Sketch Engine: the frequency wordlist gave quantitative information about the corpus, while the concordance lines helped examine the context in which the MDs occurred; the tool for annotating nodes was also adopted to tag the MDs depending on their deontic, dynamic, or epistemic meaning.

The results demonstrate the students’ preference for modals and semi-modals conveying obligation; this may be due to the topic, as the students voice their firm commitment towards their future profession. Instead, the limited use of other MDs such as lexical verbs and adverbials may depend partly on the participants’ general language proficiency in English, partly by a lack of familiarity with the conventions of academic discourse in the target language.

All in all, this study seeks to contribute to research into modality in learners’ academic writing, and the insights it offers may guide instructors of English as a foreign language in designing resources and strategies to support learners in facing the challenges that modality presents them with when developing their argumentation.

**Keywords:** learner corpora, modal devices, argumentation, subjectivity, academic writing

[9149] ***#ActuallyAutistic: ASD Community Building on Twitter.***

Savannah Brown (Oxford College of Emory University) and Jack Hardy (Oxford College of Emory University).

**Abstract.** Hashtags help connect users to content of their interest that has been made by others outside of their immediate network (Jones, 2014). They serve to create, build, and strengthen communities. Two well-known hashtags related to autism spectrum disorder (ASD) are #ActuallyAutistic and #Autism. The former is an alternative to the latter because #Autism has become dominated by organizations and non-autistic individuals (Zolyomi et al, 2020).

We were interested in exploring the origins and current discourses of tweets using each hashtag, so we created two subcorpora. One includes the first 300 tweets of #ActuallyAutistic (dating back to early 2012) and 300 tweets of #Autism that do not also include #ActuallyAutistic from the same time frame. The second subcorpus also includes 600 tweets – 300 for each hashtag – from April 2022.

Through close readings of the early #ActuallyAutistic tweets, we found that these tweets were used to establish, defend, and build the digital ASD community. We then used AntConc (Anthony, 2022) to compare wordlists and conduct keyword analyses across the subcorpora. Some notable differences we observed included the use of personal pronouns and levels of engagement. For example, #Autism tweets used more third-person pronouns, included more other hashtags, and contained more subjective adjectives (e.g., wonderful, hard). #ActuallyAutistic tweets were more interactive (using more first- and second-person pronouns and interrogatives) and contained more jargon (e.g., ABA, neurodivergent).

Although our corpus is relatively small, this study allowed us see how communities can use hashtags to find and interact with each other. We conclude our paper by suggesting potential areas of research, including “algospeak” (Day, 2022) on Autism TikTok.

**References**

Anthony, L. (2022). AntConc (Version 4.1.4) [Computer Software]. Tokyo, Japan: Waseda University. <https://www.laurenceanthony.net/software>

Day, F. (2022, January 07). Are censorship algorithms changing TikTok's culture? Medium.

Jones, J. (2014). Switching in Twitter’s hashtagged exchanges. Journal of Business and Technical Communication, 28(1), 83-108.

Zolyomi, A., Jones, R., & Kaftan, T. (2020, October). # ActuallyAutistic Sense-Making on Twitter. Conference on Computers and Accessibility.

**Keywords:** discourse analysis, autism spectrum disorder (ASD), healthcare, keyword analysis

[9196] ***Towards operationalization of the concept of discursive salience as a factor in the development of semantic prosody: the case of abortion in English Web 2020.***

Beth Malory (University College London).

**Abstract.** Notwithstanding wrangling over terminology, the concept of semantic prosody is well established, as the affective meaning a given node derives from its typical collocates (Hunston, 2007; Louw, 1993; Partington, 2004). Widely utilized in discourse studies, semantic prosody is often considered to exist on a spectrum of so-called “evaluative polarity” (Partington, 2015: 281), between extremes of positivity and negativity.

Corpus tools have been invaluable in advancing understandings of semantic prosody and associated concepts, since primings are thought to accumulate “in a quasi-statistical way”, partly determined by frequency of exposure (Partington, 2015: 293) and suited to quantitative analysis. Hoey’s (2005) theory of lexical priming, with its focus on “psychological preference” (24) and accumulative contact (8), have been central to developing a body of corpus-assisted discourse-focused research on semantic prosody. Understandings of the factors involved in the development of semantic prosody continue to evolve, however.

One potential explanatory factor yet to be explored in detail is salience. An established concept in other linguistic sub-disciplines, including sociolinguistics, psycholinguistics, and semantics (cf. Boswijk & Coler, 2020; Jaeger & Weatherholtz, 2016; Kerswill & Williams, 2002; Rácz, 2013; Zarcone et al., 2016), salience is linked to “surprisal value” (Rácz 2013: 37). Exploring its applicability in discourse analysis, this study investigates use of abortion, a highly contested lexical item, in English Web 2020. Its objective is exploring whether a node may acquire extreme semantic prosodies when associated with surprisal. The findings of this study suggest that abortion has extremely negative semantic prosody in contemporary internet English and question the contribution of extreme discourses, such as those of murder, genocide, and other forms of violence, in the development of this semantic prosody. In doing so, the reported research offers both a theoretical framework and a preliminary operationalization for discursive salience as a precursor - and contributory factor – to semantic prosody.

**References**

Boswijk, V., & Coler, M. (2020). What is Salience? Open Linguistics, 6(1), 713–722. <https://doi.org/10.1515/opli-2020-0042>

Hunston, S. (2007). Semantic prosody revisited. International Journal of Corpus Linguistics, 12(2), 249–268. <https://doi.org/10.1075/ijcl.12.2.09hun>

Jaeger, T. F., & Weatherholtz, K. (2016). What the Heck Is Salience? How Predictive Language Processing Contributes to Sociolinguistic Perception. Frontiers in Psychology, 7, 1115. <https://doi.org/10.3389/fpsyg.2016.01115>

Kerswill, P., & Williams, A. (2002). ‘Salience’ as an explanatory factor in language change: Evidence from dialect levelling in urban England. In M. C. Jones & E. Esch (Eds.), Language Change. DE GRUYTER MOUTON. <https://doi.org/10.1515/9783110892598.81>

Louw, W. E. (1993). Irony in the Text or Insincerity in the Writer? — The Diagnostic Potential of Semantic Prosodies. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), Text and Technology (p. 157). John Benjamins Publishing Company. <https://doi.org/10.1075/z.64.11lou>

Partington, A. (2004). ‘Utterly content in each other’s company’: Semantic prosody and semantic preference. International Journal of Corpus Linguistics, 9(1), 131–156. <https://doi.org/10.1075/ijcl.9.1.07par>

Partington, A. (2015). Evaluative prosody. In K. Aijmer & C. Rühlemann (Eds.), Corpus Pragmatics (pp. 279–303). Cambridge University Press. <https://books.google.co.uk/books?id=OJeiBQAAQBAJ>

Rácz, P. (2013). Salience in Sociolinguistics: A Quantitative Approach. DE GRUYTER. <https://doi.org/10.1515/9783110305395>

Stewart, D. & Routledge. (2013). Semantic prosody: A critical evaluation. Routledge.

Wiegand, V., & Mahlberg, M. (2019). Corpus Linguistics, Context and Culture. De Gruyter. <https://books.google.co.uk/books?id=OjvEDwAAQBAJ>

Yang, B., & Li, W. (2020). Corpus-based Approaches to Grammar, Media and Health Discourses: Systemic Functional and Other Perspectives. Springer Singapore. <https://books.google.co.uk/books?id=Kf7sDwAAQBAJ>

Zarcone, A., van Schijndel, M., Vogels, J., & Demberg, V. (2016). Salience and Attention in Surprisal-Based Accounts of Language Processing. Frontiers in Psychology, 7. <https://doi.org/10.3389/fpsyg.2016.00844>

**Keywords:** Semantic prosody, Discourse, Salience, Lexical priming, Evaluation, Reproductive rights, Collocation

[9209] ***The language of children's online disclosures of abuse.***

Mark McGlashan (Birmingham City University).

**Abstract.** Significant barriers exist for children and young people when disclosing experiences of abuse. A known issue is that children are more likely to disclose to same-aged peers rather than to a close or familiar adult (Howell et al., 2014) as children may distrust how adults might respond to their disclosures (Callaghan et al., 2017).

As such, problems exist for both children whose voices may be obscured/silenced but also for practitioners and researchers who have limited access to authentic, empirical disclosures of abuse made by children that could enable practitioners to more effectively respond to children’s concerns during disclosure.

Childline is a free public service providing support for children in the UK and provides online message boards (<https://www.childline.org.uk/get-support/message-boards/>) as a venue for children to anonymously seek support for a range of issues from other anonymous peers using the site. Posts seeking support often also include children disclosing experiences of abuse and there are boards specifically designated for support relating to emotional, physical and sexual abuse. As such, boards such as Childline offer the potential to capture children’s reports – in their own words – of abuse, and enable us to address the central question posed by this paper: what language is used by children when they make disclosures of abuse?

Specifically, the paper reports on two separate studies conducted using a corpus of 3,242 posts (~520,000 tokens) scraped from Childline’s message boards which focus on: (i) children’s reports of domestic abuse; and (ii) comparing the language in reports of emotional, physical, and sexual abuse. Findings from this work may help provide a better empirical linguistic understanding of children’s experiences of abuse as disclosed in an online, self-selecting, asynchronous forum. Such insight may be of direct, practical use to relevant practitioners as a way to facilitate more effective disclosures of abuse.

**References**

Callaghan, J., Fellin, L., Mavrou, S., Alexander, J. & Sixsmith, J. (2017) The Management of Disclosure in Children’s Accounts of Domestic Violence: Practices of telling and not telling, Journal of Child and Family Studies, 26, 3370—3387.

Howell, K. H., Cater, Å. K., Miller-Graff, L. E., & Graham-Bermann, S. A. (2014) The process of reporting and receiving support following exposure to intimate partner violence during childhood, Journal of Interpersonal Violence, 30, 2886-2907.

**Keywords:** Corpus Linguistics, Forensic Linguistics, Abuse, Disclosure

[9243] ***Exploring the weakening of 'fuck' in casual conversation.***

Robbie Love (Aston University) and Anna-Brita Stenström (University of Bergen).

**Abstract.** In contemporary British English, FUCK is a very common swear word, despite its long-standing reputation as one of the strongest ones. This study examines the use of FUCK among British English teenagers in casual conversation, by comparing youth language corpora from two different periods, the Bergen Corpus of London Teenage Language (COLT), recorded in 1992, and a southeast English sub-corpus of teenage speech from the Spoken British National Corpus 2014 (Spoken BNC2014).

The study takes a pragmatic approach to the use of swear words, focussing on their emphatic and potentially offensive roles in dialogic communication. We explore how the forms and functions of FUCK may have changed among teenagers in casual conversation between the 1990s and 2010s, with a view to establishing whether the empirical data can any longer support an interpretation of FUCK as a ‘strong’ swear word.

In coding all instances of FUCK in both corpora using an adapted version of McEnery & Xiao’s (2004) ‘categories of insult’ scheme, we find evidence of semantic broadening indicated by increased usage of FUCK in idiomatic (fixed) expressions, e.g. ‘what the fuck’; a greater presence of figurative and general expletive usage; and general emphatic usage.

Furthermore, we find very little evidence of FUCK being used as a potential term of abuse. We conclude that contemporary usage of FUCK by young speakers is resemblant of a relatively mild swear word, offering further resolution of the ‘swearing paradox’ (Beers-Fägersten, 2007), whereby swearing frequency and offensiveness ought to be negatively correlated.

**References**

Beers Fägersten, K. 2007. A sociolinguistic analysis of swearword offensiveness. Saarland Working Papers in Linguistics, 1, 14-37.

McEnery, T., & Xiao, R. 2004. British English: the case of FUCK in the BNC. Language and Literature, 13(3), 325-368.

**Keywords:** swearing, teenage language, casual conversation, swearing paradox, semantic weakening

[9245] ***How reliable is our reference? The effect of the reference corpus on measuring lexical sophistication in L2 English speech.***

Raffaella Bottini (Lancaster University) and Elen Le Foll (UCLouvain).

**Abstract.** Measures of lexical sophistication quantify “the proportion of relatively unusual or advanced words” in a text, “appropriate to the topic and style” of the communicative event (Read, 2000:200-203). Mean-frequency metrics, in particular, compute the average frequency value of a text by extracting a frequency score for each of its words using a reference corpus: the lower the mean frequency score of a text, the more sophisticated its vocabulary. Brysbaert and New (2009) argue that the choice of a reference corpus that matches the register of the target texts to be analysed is crucial, and Tidball and Treffers-Daller (2008) suggest the use of spoken reference corpora in research on L2 speech. Similarly, Egbert (2017) highlights the importance of a reference corpus’ situational variables. However, to date, most studies on L2 spoken production have relied on a mixture of written and spoken corpora, without matching the mode and/or register of the reference corpora to the target texts. In addition, spoken reference corpora often consist of scripted language. The fact that the representativeness of the reference corpus and its comparability to the target texts are rarely evaluated raises serious validity and generalisability issues.

In this study, we compute mean-frequency scores of lexical sophistication in L2 English speech comparing different reference corpora to investigate the effect of register on lexical sophistication scores. To this end, we rely on the ICNALE corpus, a unique dataset of over 3 million words of L2 English (ranging from A2 to B2 CEFR proficiency levels) with which we can control for the potential effects of topic and production time (Ishikawa, 2013; 2014). We argue that the use of a reference corpus that does not match the learner production mode, or combines different registers and/or language modes, negatively impacts the validity of mean-frequency scores of lexical sophistication.

**References**

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. Behavior Research Methods, 41, 977–990.

Egbert, J. (2017). Corpus linguistics and language testing: Navigating uncharted waters. Language Testing, 34(4), 555–564.

Ishikawa, S. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. Learner corpus studies in Asia and the world, 1, 91–118.

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.

Read, J. (2000). Assessing vocabulary. Cambridge University Press.

Tidball, F., & Treffers-Daller, J. (2008). Analysing lexical richness in French learner language: What frequency lists and teacher judgement can tell us about basic and advanced words. French Language Studies, 18(3), 299–313.

**Keywords:** lexical sophistication, reference corpus, L2 English speech

[9327] ***A Corpus-Assisted Analysis of Shell Noun Use in Academic Lectures.***

Haoshan Ren (Georgia State University) and J. Elliott Casal (The University of Memphis).

**Abstract.** Shell nouns (SN) have been extensively studied in corpus-based research on academic writing practices (e.g., Aktas & Cortes, 2008; Gray & Cortes, 2011) for their function to reference and label complex ideas in surrounding discourse (Schmid, 2000; also ‘signaling noun’ by Flowerdew, 2003). However, despite the important role SNs play in organizing information in academic lectures (Flowerdew & Forest, 2015), relatively less attention has been paid to SNs in this genre (Benitez-Castro & Thompson, 2015).

Therefore, our project analyzes SN use in university lectures with particular attention to their structures and functions in questions, and in conjunction with importance/relevance markers (Chang, 2012; Deroey, 2012; Deroey & Taverniers, 2012a, 2012b). This presentation aims to provide pedagogical implications and complement previous literature that highlights SNs’ referential nature and structural prevalence in academic lectures.

Our data consists of 30 complete lecture transcripts (320,795 words) sampled from MICASE (Simpson, Briggs, Ovens, & Swales, 2002), balanced evenly across Physical Sciences and Social Sciences academic divisions. Lectures were part-of-speech tagged, and SN candidates were identified via python scripts based on Flowerdew and Forest’s (2015) signaling noun list and Schmid’s (2000) SN list. We then manually reviewed each tagged SN. The confirmed instances were then classified according to Flowerdew and Forest’s (2015) semantic taxonomy.

This presentation is part of an ongoing project that analyzes SNs in academic lectures in terms of their frequency, semantic category distribution, type frequency ranks, and the functional patterns of SN use in the lectures overall and across academic divisions. Our results indicate that SN plays a significant role in questions and importance/relevance markers in academic lectures. The frequency of SN use across disciplines highlights the important role they play in establishing cohesion in lectures, while the SN type distribution reveals the idiosyncratic nature of SN types in academic spoken discourse.

**Keywords:** Shell nouns, signaling nouns, academic lectures, textual variation across disciplines and interactivity levels

[9470] ***A corpus-based critical discourse analysis of language ideologies in parliamentary debates about the recognition of Irish Sign Language.***

Robyn Cunneen (University of Limerick) and Maria Rieder (University of Limerick).

**Abstract.** ISL (Irish Sign Language) is the national sign language of Ireland, used by approximately 40,000 people, 5,000 of which are deaf (Irish Deaf Society, 2023). ISL became an officially recognised language in Ireland by means of the ISL Act 2017, which commenced in December 2020.

This came about after more than 30 years of campaigning by the Irish Deaf Society and included social media and census campaigns. Prior to this legislation, ISL was only mentioned in the Education Act 1998, being referred to as a “support service”. While some work has investigated language ideologies behind the ISL recognition campaign such as Conama (2020), this study explores language ideologies in parliamentary discourse, specifically perspectives of languageness of ISL. This is crucial to the study of sign language recognition and policymaking, as differing views on sign languages and deafness can lead to different policy outcomes (Reagan, 2019).

A corpus of parliamentary debates on ISL recognition (2013-2020) was compiled and analysed using a combination of corpus linguistics tools (Sketch Engine) and critical discourse analysis (CDA). Specifically, Van Dijk’s sociocognitive approach, along with Van Leeuwen’s (2007, p. 91) framework for analysing “the way discourses construct legitimation for social practices” will be drawn upon. Theoretically, Krausneker’s (2015) categories of sign language ideologies will be applied.

Initial findings point to a major theme of language legitimacy. Specifically, comparisons to the Irish language (Gaeilge) as a minority language of the State are made, claims about the nature of ISL with respect to it being a “native” or “indigenous” language of the deaf community are argued and benefits of recognising ISL on wider society are highlighted.

**References**

Conama, J.B. (2020). 35 Years and Counting! An Ethnographic Analysis of Sign Language Ideologies within the Irish Sign Language Recognition Campaign. In A. Kusters, M. Green, E. Moriarty and K. Snoddon (Eds.), Sign Language Ideologies in Practice (pp. 265-286). De Gruyter.

Irish Deaf Society. (2023). Irish Sign Language. <https://www.irishdeafsociety.ie/irish-sign-language/>

Krausneker, V. (2015). Ideologies and Attitudes toward Sign Languages: An Approximation. Sign Language Studies, 15(4), 411–431. <http://www.jstor.org/stable/26190996>

Reagan, T. (2019). Language policies, language rights, and sign languages: A critique of disability-based approaches. Critical Inquiry in Language Studies, 16(4), 271-292. <https://doi.org/10.1080/15427587.2019.1574577>

Van Leeuwen, T. (2007). Legitimation in discourse and communication. Discourse & Communication, 1(1), 91–112. <https://doi.org/10.1177/1750481307071986>

**Keywords:** Critical Discourse Analysis, Corpus Linguistics, Sign Language Ideologies, Irish Sign Language

[9510] ***Exploring individual variation in constructional schematicity using random effects.***

Svetlana Vetchinnikova (University of Helsinki).

**Abstract.** This paper builds on two theses: 1) language is an inventory of constructions at different levels of schematicity and 2) personal construct-i-cons vary as a function of usage. With repeated use, constructions move along the continuum from more schematic to more lexically specified through the process of chunking and can undergo reanalysis. 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 typical property of chunks and can serve as a diagnostic of a change in the internal structure of an expression. Given individuality of language usage, to what extent do different instantiations of constructions vary in schematicity in personal construct-i-cons?

As a case study, I used a 1.75-million-word corpus of comments posted by one blogger over 8 years. As a dependent variable, I chose the alternation between contracted and uncontracted forms of it is hypothesizing that it was more likely to be reduced in chunks. It is occurs in a wide variety of syntactic structures including clefts, progressives, passives, extraposed and copular structures: altogether 10,000 corpus occurrences of it is/it’s were categorized into 15 frequent constructions.

For each lexical item filling the open slot, I used delta P statistic to compute the degree to which it associates with a construction and the degree to which the construction associates with it (Gries & Ellis 2015). In addition, in a logistic regression model predicting the contracted form, I included possible priming and temporal order of occurrence as fixed effects and lexically specified instantiations of constructions as random effects.

Variance in random intercepts showed variation of lexically specified instantiations in schematicity and variance in random slopes for the effect of temporal order showed change in schematicity over time.

**References**

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. <https://doi.org/10.1515/ling.37.4.575>.

Gries, Stefan Th & Nick C. Ellis. 2015. Statistical measures for usage-based linguistics. Language Learning 65(S1). 228–255. <https://doi.org/10.1111/lang.12119>.

**Keywords:** individual differences, chunking, constructions, reduction, mixed models

[9513] ***A multidimensional text typology of American film.***

Marcia Veirano Pinto (Federal University of Sao Paulo (UNIFESP)) and Tony Berber Sardinha (Pontifical Catholic University of São Paulo).

**Abstract.** The aim of this paper is to present the first linguistic typology of American movies derived from a multi-dimensional (MD) analysis, which was based on an extended version of the NAMC - North American Movie Corpus (Veirano Pinto, 2014). The extended NAMC comprises the full dialogue of 720 movies from four different genres, namely action/adventure, comedy, drama, and horror/suspense/mystery released between 1930 and 2020. The corpus includes 180 movies of each genre, 80 per decade. The text types presented in this study are based on the dimension scores of each movie on the seven dimensions of variation identified, which were entered into a stepwise cluster analysis performed using PROC FASTCLUS in SAS University Edition. The steps followed in the identification of the text types were: (1) Enter the dimension scores into a stepwise cluster analysis; (2) Identify the number of clusters in the data; (3) Extract the clusters; (4) Compute the cluster means distance; (5) Use cluster means to identify core (texts that are prototypical of the cluster because they include more of the linguistic features loading on the dimension) and peripheral (texts that have fewer of these features) texts; (6) Check the distribution of texts across the clusters; (7) Describe the dimensional profile of the clusters according to their mean dimension scores; (8) Interpret the clusters communicatively to determine the text types.

The three clusters that emerged were labeled as Planned Dialogue with a Situational Focus (text type 1), Spontaneous Articulation of Verbal Events (text type 2), and Stance Marked Argumentative Dialogue (text type 3). The movies that entered in each of these clusters reveal distinct linguistic trends across time and genres.

**Keywords:** text types, american movies, multi-dimensional analysis

[9530] ***From specialized corpus to the EAP classroom: Integrating authentic data into materials design.***

Ana Luiza Pires de Freitas (Federal University of Health Sciences of Porto Alegre/ Brazil), Ana Eliza Pereira Bocorny (Federal University of Rio Grande do Sul/ Brazil), Rozane Rodrigues Rebechi (Federal University of Rio Grande do Sul/ Brazil) and Simone Sarmento (UFRGS).

**Abstract.** This paper presents a roadmap for extracting and integrating corpus data into materials designed for English for Academic Purposes (EAP) writing courses. It stems from the recognition that, even though in the early 2000s Sinclair (2004) anticipated that corpus-based language learning and teaching would revolutionize language pedagogy, and many researchers emphasize the importance of integrating corpus data into EAP pedagogy (Flowerdew 2009, 2013, 2014; Gray et al. 2020; Charles & Frankenberg-Garcia 2021), few EAP teachers actually know how to extract and utilize such data in their practice (Frankenberg-Garcia et al. 2020; Römer 2006). Having this said, the aim of this paper is twofold: (i) help EAP teachers better understand corpus linguistics methods for the extraction of language data from specialized corpora and (ii) show how said language data can be used in the design of EAP writing course materials through a pedagogical model that combines corpus and genre-based approaches.

To reach the proposed goals, first, we review corpus and genre-based approaches to language learning and teaching and pedagogical models that combine both approaches (Charles, 2007; Cotos et al. 2017; Flowerdew, 2012; Gray et al. 2020, McEnery and Hardie, 2012; Sinclair, 1991). After that, we describe a framework and principles for designing EAP materials that combine corpus and genre-based pedagogies (Bocorny & Welp, 2021; Moreno and Swales, 2018; Reppen, 2010; Schneuwly and Dolz, 2004). Next, to support the design of a pedagogical unit according to the framework suggested, we introduce guidelines based on principles suggested by Welp et al. (2019) and adapted by Bocorny and Welp (2021), as a checklist to be followed. Finally, we highlight the feasibility of such a model for novice and experienced teachers who can use the step-by-step guide combining corpus and genre-based approaches for academic writing.

**References**

Bocorny, A. E. P., & Welp, A. K. D. S. (2021). O desenho de tarefas pedagógicas para o ensino de inglês para fins acadêmicos: conquistas e desafios da Linguística de Corpus. Revista de estudos da linguagem. Belo Horizonte, MG. Vol. 29, n. 2 (abr./jun. 2021), p.[1589]-1638.

Charles, M., & Frankenberg-Garcia, A. (2021). Introduction: Dichotomies and debates in corpora and ESP/EAP writing. In Corpora in ESP/EAP Writing Instruction (pp. 1-10). Routledge.

Charles, M. (2007). Reconciling top-down and bottom-up approaches to graduate writing: Using a corpus to teach rhetorical functions. Journal of English for academic purposes, 6(4), 289-302.

Cotos, E., Haufman, S. & Link, S. (2017). A move/step model for methods sections: Demonstrating Rigour and Credibility. English for Speficic Purposes. 46, 90-106.

Flowerdew, L.(2014). Corpus-based analyses in EAP. In Academic discourse (pp. 105-124). Routledge.

\_\_\_\_\_\_\_\_\_\_\_\_\_ (2013). Corpus-based research and pedagogy in EAP: From lexis to genre. Language Teaching, 48(1), 99-116.

\_\_\_\_\_\_\_\_\_\_\_\_\_ (2012). Corpus and Language Education. Basingstoke: Palgrave Macmillan.

\_\_\_\_\_\_\_\_\_\_\_\_\_ (2009). Applying corpus linguistics to pedagogy: A critical evaluation. International journal of corpus linguistics, 14(3), 393-417.

Frankenberg-Garcia, A., Rees, G. P., & Lew, R. (2020). Slipping Through the Cracks in e-Lexicography. International Journal of Lexicography, 34(2), 206-234.

Gray, B., Cotos, E., & Smith, J. (2020). Combining rhetorical move analysis with multi-dimensional analysis: Research writing across disciplines. Advances in corpus-based research on academic writing: Effects of discipline, register, and writer expertise, 137-168.

McEnery, T. & Hardie, A. (2012). Corpus Lingusitics: Method, Theory and Practice. Cambrdge: Cambridge University Press.

Moreno, A. I., & Swales, J. M. (2018). Strengthening move analysis methodology towards bridging the function-form gap. English for Specific Purposes, 50, 40-63.

Reppen, R. (2010). Using corpora in the language classroom. Cambridge University Press.

Römer, U. (2006). Pedagogical Applications of Corpora: Some Reflections on the Current Scope and a Wish List for Future Developments. Zeitschrift für Anglistik und Amerikanistik, 54(2), 121-134. <https://doi.org/10.1515/zaa-2006-0204>

Schneuwly, J. & Dolz, B. (2004). Gêneros orais e escritos na escola: Mercado de Letras, Campinas.

Sinclair, J. (2004). Trust the text: Language, Corpus and Discourse. London/ New York. Routledge.

\_\_\_\_\_\_\_\_\_ (1991). Corpus, Concordance, Collocation. Oxford: OUP.

Welp, A. K., Didio, Á. R., & Finkler, B. (2019). Questões contemporâneas no cinema e na literatura: o desenho de uma sequência didática para o ensino de inglês como língua adicional. BELT-Brazilian English Language Teaching Journal, 10(2), e35861-e35861.

**Keywords:** English for Academic Purposes, Corpus-Based Pedagogy, EAP Material Design

[9631] ***Structure signalling in L2 undergraduate academic writing: A cross-discipline comparison.***

Linfeng Cai (Hong Kong Baptist University) and Meilin Chen (Hong Kong Baptist University).

**Abstract.** Theorising an array of lexical devices for writers to interact with readers and navigate them through discourses, the metadiscourse model (Hyland, 2005) has sparked massive research interest, mainly in professional research writing (Adel, 2022; Cao & Hu, 2014; Hyland & Jiang, 2020). When it comes to the smaller number of studies on undergraduate writing, the common emphasis is on students’ establishment of stances (Aull & Lancaster, 2014; Crosthwaite & Jiang, 2017; Yoon & Romer, 2020) rather than the structuring of text. While some research has suggested students’ customary labelling of the concluding stage and the increased tendency of high-performing writers to disclose their argumentative intentions (Basturkmen & von Randow, 2014; Ho & Li, 2018), focused investigations into structure construction with frame markers are lacking. The role of section headings in structuring texts has been especially overlooked, except for a few (Gardner & Holmes, 2009, 2010).

To fill the research gap, this study explores how four types of frame markers, i.e., headings, goal announcers, stage labellers, and topic shifts, are utilised in undergraduate subject course assignments in three disciplines: Chinese Medicine, Education, and Sociology.

For each discipline, a corpus of roughly 30 to 60 English assignments (50,000 to 130,000 words) by L2 learners was built and manually tagged. Preliminary results show both differences and commonalities across the disciplines. Compared with writing in the other two disciplines, Chinese Medicine assignments contain fewer announcements of discourse goals on average and are partitioned by section headings less frequently. In general, student writers tend to mark the concluding stage explicitly but denote topic shifts or outline texts infrequently. When goals are announced, this is usually realised by a limited range of lexical expressions. Overuse of specific devices, particularly headings, has also been observed in several cases. These findings imply a need for explicit instruction on adequate and appropriate use of frame markers and an expansion of students’ repertoires to help them produce more coherent and convincing academic proses in their own disciplinary communities.

**References**

Adel, A. (2022). Writer and reader visibility in humanities research articles: Variation across language, regional variety and discipline. English for Specific Purposes, 65, 49-62.

Aull, L. L., & Lancaster, Z. (2014). Linguistic markers of stance in early and advanced academic writing: A corpus-based comparison. Written Communication, 31(2), 151-183.

Basturkmen, H., & von Randow, J. (2014). Guiding the reader (or not) to re-create coherence: Observations on postgraduate student writing in an academic argumentative writing task. Journal of English for Academic Purposes, 16, 14-22.

Cao, F., & Hu, G. W. (2014). Interactive metadiscourse in research articles: A comparative study of paradigmatic and disciplinary influences. Journal of Pragmatics, 66, 15-31.

Crosthwaite, P., & Jiang, K. (2017). Does EAP affect written L2 academic stance? A longitudinal learner corpus study. System, 69, 92-107.

Gardner, S., & Holmes, J. (2009). Can I use headings in my essay? Section headings, macrostructures and genre families in the BAWE corpus of student writing. In M. Charles, D. Pecorari, & S. Hunston (Eds.), Academic writing : At the interface of corpus and discourse (pp. 251-271). London: Continuum.

Gardner, S., & Holmes, J. (2010). From section headings to assignment macrostructures in undergraduate student writing. In E. Swain (Ed.), Thresholds and potentialities of systemic functional linguistics: Multilingual, multimodal and other specialised discourses (pp. 268-290). Trieste, Italy: EUT Edizioni Università di Trieste.

Ho, V., & Li, C. (2018). The use of metadiscourse and persuasion: An analysis of first year university students’ timed argumentative essays. Journal of English for Academic Purposes, 33, 53-68.

Hyland, K. (2005). Metadiscourse: Exploring interaction in writing. London: Continuum.

Hyland, K., & Jiang, F. (2020). Text-organizing metadiscourse tracking changes in rhetorical persuasion. Journal of Historical Pragmatics, 21(1), 137-164.

Yoon, H. J., & Romer, U. (2020). Quantifying disciplinary voices: An automated approach to interactional metadiscourse in successful student writing. Written Communication, 37(2), 208-244.

**Keywords:** academic writing, metadiscourse, learner corpora, English for academic purposes

[9642] ***Linguistic and Situational Granularity within Registers: A Cross-register Perspective.***

Marianna Gracheva (Northern Arizona University) and Jesse Egbert (Northern Arizona University).

**Abstract.** Registers are text varieties associated with particular situational factors (e.g., participants’ roles, relationships, goals, setting, mode) (Biber, 1988). It has also been shown that situational variation exists not just between registers, but also within registers, and that situational variation within registers is functionally related to linguistic differences (e.g., Biber, 1994; Gray, 2015; Biber et al., 2021; Egbert & Gracheva, 2022).

Thus, the functional correspondence between situation and language can be traced at increasingly granular levels of analysis even within culturally-recognized registers. Crucially, this body of research demonstrates that the granular approach to situation can account for functionally interpretable linguistic variation previously unaccounted for by existing register categories.

The present study extends this research through a comparative cross-register investigation of the degree of internal variation permitted within seven registers: emails, text messages, image descriptions, evaluations, interviews, business memos, and essays. The study focuses on communicative purpose as a situational parameter, and each text is coded for presence or absence of specific communicative purposes. Purpose is then examined as a predictor of intra-register variation on the linguistic dimensions of variation identified through a multidimensional analysis. The study addresses the following questions:

1. To what extent does communicative purpose predict linguistic variation within each register?

2. How do registers compare in terms of the internal situational and linguistic variation they allow?

Thus, the study first identifies the full range of communicative purposes fulfilled by the analyzed registers. The results demonstrate that communicative purpose is a major predictor of increasingly granular linguistic variation within registers. However, importantly, our findings also show that registers differ in (a) the extent to which they rely on the identified purposes; (b) the amount of register-internal linguistic freedom accounted for by purpose.

**Keywords:** within-register variation, situational granularity, linguistic granularity, communicative purpose, functional linguistic variation, functional correspondence

[9785] ***A new method for coding multimodal corpora: the discursive construction of free-range animals on Twitter.***

Yuze Sha (Lancaster University) and Beth Malory (University College London).

**Abstract.** Language-only corpora place restrictions on the scope of corpus studies about communication due to the richness of multimodal environments, such as social media platforms, which many experience daily. Here, visual and auditory resources may be just as important as text, but many recent large-scale quantitative discourse studies would overlook these. This is, in large part, because the interactivity of semiotic modes poses methodological challenges for corpus linguists. Qualitative discourse studies, involving small samples of texts, circumvent this issue by exploring inter-mode meaning-making processes with specific examples (cf. Brookes et al., 2018), but scaling this up to larger datasets is challenging.

Alternative approaches proposed in recent years, for example, Collins (2020) and Christiansen et al. (2020) have suggested methodologies applicable only in restricted contexts (e.g., emojis in Collins, 2020), or required ‘translation’ of other semiotic modes into text (e.g., Christiansen et al., 2020), risking meaning change.

This research introduces a novel method for developing and coding multimodal corpora with Atlas.ti, software originally designed for qualitative data analyses. Presenting a case study of Twitter representations of free-range animals (hereafter FRA), we demonstrated the utility of Atlas.ti’s Code Co-occurrence Table function in annotating multimodal concordances and in the present case, comparing different identity groups’ construction of their relationships to FRA. Adapting van Leeuwen’s (2008) social actor network and Martin and White’s Appraisal Framework’s Attitude dimension (2003), this study revealed the division on Twitter that animal activists overwhelmingly construct animals as beings (independent from consumption), whereas commercial representations were of animal things (food and products). These patterns were analysed combining their textual and visual dimensions, with presentations of “flocks” correlating to negative attitudes and clustering in activist discourses, and those of “few” animals correlating to positive attitudes and commercial discourses. The construction and coding of this multimodal corpus, therefore, allowed the triangulation of frequency, collocation, and concordance analysis for the in-depth investigation of this multimodal dataset.

**Keywords:** multimodal corpus, corpus-assisted multimodal critical discourse studies, social media, corpus annotation

[9876] ***Can they learn from their own mistakes? An investigation into the application of learner corpora in an English academic writing course for L2 undergraduate students.***

Meilin Chen (Hong Kong Baptist University).

**Abstract.** The corpus-assisted approach to language teaching, i.e., getting learners to consult corpora in or out of the classroom, has been found to be effective in expanding learners’ lexical knowledge (e.g., Frankenberg-Garcia, 2019), improving their writing via corpus-assisted self-corrections (e.g., McGarrell, 2015), and fostering greater awareness of disciplinary academic writing conventions (e.g., Cortes, 2014; see Boulton & Cobb, 2017; Chen & Flowerdew, 2018 for reviews). However, as many scholars have pointed out (Boulton & Vyatkina, 2021; Chen & Flowerdew, 2018; L. Flowerdew, 2015; Granger, 2009; Granger et al., 2015; Huang, 2017; McEnery et al, 2019), the application of learner corpora in the language classroom has been scant. To further the discussion about the potential of learner corpora in facilitating language learning, this study investigates the benefits as well as the limitations of using a learner corpus in a discipline-specific academic writing course.

The learner corpus utilized in this study consists of English written assignments by L2 learners from three disciplines, i.e., Education, Sociology, and Chinese Medicine, for their subject courses. Prior to the course, analyses of the corpus were carried out to identify problematic issues in learner writing. The teaching materials were then developed based on the results, focusing on interactive metadiscourse devices (Hyland, 2004), including frame markers (e.g., this paper aims to, in summary), transition makers (e.g., however, therefore), and endophoric markers (e.g., as mentioned previously). During the four-session course, printed concordancing lines containing these devices from each disciplinary sub-corpus, together with those from an expert corpus that was comprised of academic journal articles, were given to the experimental group. The control group, on the other hand, received materials from the expert corpus only. Multiple types of data were collected, including weekly reflective journals, post-course group interviews, email interviews, pre- and post-course writing samples.

The findings show that, while the learners in the experimental group acknowledged the benefits of comparing examples from both learner and expert writing in their weekly reflective journals, they valued the “models” from the expert writing more in the post-course interviews. This may be partly due to the fact that most learners are lower-level undergraduate students, who are eager for “good models” that they could follow when they write.

The analysis of students’ pre- and post-course writing reveals that in their post-course writing some students from the experimental group still made the same mistakes that they already identified in the in-class concordancing materials, even though they claimed in the interviews and reflective journals that they tried to incorporate what they had learned in the class into their writing. Discussions about why the use of learner corpora did not make a strong impact on learners’ writing development will be given, which provide valuable pedagogical implications for how learner corpora could be better integrated into the English academic writing classroom.

**Keywords:** academic writing, data-driven learning, learner corpora, corpus-assisted teaching, writing across disciplines, discipline-specific writing

[9944] ***A corpus-driven study of proxies for register levelling in Present-Day British English.***

Laura Abalo-Dieste (Universidade de Vigo).

**Abstract.** Writing is characterised by an overuse of informational structures and an avoidance of involved devices (Biber 1988), but studies on sociolinguistic phenomena such as colloquialisation (Mair 1997) have revealed an uncharacteristic growth of features typical of spoken language (e.g. contractions) in formal writing. The rationale behind this variation is problematic since these linguistic features can be accounted for by one or more sociolinguistic processes (Baker 2017).

This study explores twenty linguistic features that previous research took as evidence of these sociolinguistic phenomena. The aim is two-fold: to develop a theoretical framework that could account for the stylistic shift described in the literature, and to compile a complete inventory of linguistic choices contributing to it. To that aim, diachronic and stylistic data were retrieved from the BNC1994 (BNC Consortium 2007) and the BNC2014 (Love et al. 2017; Brezina et al. 2021) using #LancsBoxX (Brezina & Platt 2022).

The frequency rates in speech and writing of sixteen of the features analysed have significantly narrowed in the time span analysed, a process I refer to as ‘register levelling’. The patterns include: first, features typical of speech which are increasing in writing but decreasing in speech; second, features typical of speech which are increasing in both registers; and third, features typical of writing which follow the opposite patterns, i.e. a decrease in writing only or in both registers. The other four features show divergence between speech and writing, i.e. their preference in speech or writing is accentuated over time. Standardized residuals confirmed that features typical of speech are overused in fiction, elanguage, formal and informal speech, in contrast with features typical of writing which are overused in academic prose, official documents, magazine, newspapers, and written-to-be-spoken. To further explore the relation between register levelling and formality, a correspondence analysis will be applied.

**References**

Baker, Paul. 2017. American and British English: Divided by a Common Language? Cambridge, UK: Cambridge University Press.

Biber, Douglas. 1988. Variation across Speech and Writing. Cambridge: Cambridge University Press.

BNC Consortium. 2007. British National Corpus: XML edition. Oxford Text Archive.

Brezina, Vaclav, Abi Hawtin & Tony McEnery. 2021. The written British National Corpus 2014 – design and comparability. Text & Talk 41(5–6), 595–615.

Brezina, Vaclav & William Platt. 2022. #LancsBox X. [software].

Love, Robbie, Claire Dembry, Andrew Hardie, Vaclav Brezina & Tony McEnery. 2017. The Spoken BNC2014: designing and building a spoken corpus of everyday conversations. International Journal of Corpus Linguistics, 319–344.

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, 195–209. Amsterdam: Rodopi.

**Keywords:** diachronic, stylistic change, Present-Day English

# Abstracts: Posters

[1] ***Corpus Analysis Tools: A Resource Guide***

Martin Wynne (University of Oxford).

**Abstract.** As part of the CLARIN Resource Families initiative, a new guide to desktop and online corpus analysis tools has been written and will be published online in early 2023. The guide provides information for users on how to find, select and get started with software applications for corpus analysis. The guide aims to cover all applications with user interfaces that, as a minimum, can offer functionality to access texts and language corpora and carry out at least basic operations such as concordancing. The guide will be kept up to date, and published on the CLARIN website.

Registries of information about software are notoriously out of date. By being embedded in the CLARIN infrastructure with a long-term perspective and the backing of national infrastructures and funders, it is hoped that the CLARIN Resource Family for Desktop and Online Corpus Analysis Tools will have a better chance of long-term sustainability.

**Keywords:** Corpus analysis tools, Registry, Resource discovery, Concordancing

[2] ***Wordless: An Integrated Corpus Tool with Multilingual Support for the Study of Language, Literature, and Translation.***

Lei Ye (Shanghai International Studies University).

**Abstract.** Wordless (https://github.com/BLKSerene/Wordless) is an integrated corpus tool with multilingual support designed to be a more powerful alternative to other commonly-used corpus tools like WordSmith, AntConc, Sketch Engine, ParaConc, etc.

Wordless is battery-included with a complete set of industrial-strength NLP tools built on bleeding-edge technologies and state-of-the-art research for 100+ languages with all implementation details hided to free end users from tricky technical particularities and hence facilitate wholehearted dedication to research/studies.

Wordless has implemented a much greater variety of measures used to calculate dispersion, adjusted frequency, statistical significance, Bayes factor, effect size, etc., and supports a broader range of data visualization options including dispersion plots, line charts, word clouds, network graphs, and dependency graphs.

Wordless features an intuitively designed graphical interface with Unicode and localization support particularly geared toward non-technical users. It is open-source and freely available to anyone for any purpose without registration/subscription and equipped with full cross-platform support.

**Keywords:** corpus tool, multilingual support, battery-included, statistical measures, data visualization, non-technical user, open-source

[3] ***Introducing CONSOS: A Spoken Corpus of New Speakers of Scots.***

Linda Bruce (The Open University).

**Abstract.** With 1.5 million speakers, Scots is Scotland’s largest minority language. Research on Scots tends to focus on L1 speakers, while less scholarly attention has been paid to new speakers (i.e., adults with no background in the language, who choose to learn and use it habitually). However, work on minority languages including Welsh, Irish and Scottish Gaelic (Hornsby & Vigers, 2018; McLeod, 2018; O'Rourke & Walsh, 2020) acknowledges new speakers as a social group worth investigating because they share common characteristics, ideological positions, and language practices. Compiled from interview data, CONSOS is the first spoken corpus of new speakers of Scots. An explanation of how the corpus was constructed will be provided alongside initial findings concerning morphosyntactic variation across new and L1 speakers. Preliminary data on the ideologies indexed and identity affiliations adopted by new speakers will also be presented.

**References**

Hornsby, M., & Vigers, D. (2018). ‘New’speakers in the heartlands: struggles for speaker legitimacy in Wales. Journal of multilingual and multicultural development, 39(5), 419-430.

McLeod, W. (2018). New speakers of Gaelic: A historical and policy perspective. In Gaelic in Contemporary Scotland: The Revitalisation of an Endangered Language (pp. 79-93). Edinburgh University Press.

O'Rourke, B., & Walsh, J. (2020). New Speakers of Irish in the Global Context: New Revival? (1 ed.). Routledge.

**Keywords:** Scots, New Speakers, Minority Languages, Morphosyntactic Variation, Corpus Construction

[4] ***Texts of the Scottish Reformation: Creating and Normalising Corpora.***

Beth Beattie (University of Glasgow).

**Abstract.** Exploring the linguistic realisation of the Scottish and English Reformations can tell us more about the expression of nascent British national identities. Comparing the lexical choices of sixteenth-century Reformed and Catholic writers illustrates how religious communities differentiated themselves across the British Isles. There is little surviving textual evidence from Catholic sources in Scotland (Ryrie 2004), especially when compared with prolific Reformed writers like John Knox and David Lindsay. So, how can Reformed and Catholic sources be compared in a representative way? This paper discusses the benefits of moving from a broad-scale comprehensive study to a series of focused case studies to ensure a greater comparability of sources. Furthermore, what might be required to adapt spelling normalisation software, such as VARD, to Older Scots? Building on the work of Kopaczyk (2013), this paper also outlines specialised Older Scots training developed for VARD, which has broader applications beyond religious texts.

**Keywords:** Early Modern, corpus design, spelling normalisation, Older Scots

[5] ***Analysing the Circulation of Endometriosis Terms in Different Speech Communities: Challenges for the corpus compilation.***

Julie Humbert-Droz (CeRLA, University of Lyon 2).

**Abstract.** The circulation of medical terms in different speech communities raises important issues of understanding and appropriation of terms by laypeople and patients (e.g. Gill & Maynard 2006, León-Araúz 2015, Delavigne et al. 2022). This poster will address challenges related to the building of a corpus intended to represent the ways endometriosis terms are used by experts, patients, and laypeople. Compiling this corpus is in fact the first stage of a project that aims to study the circulation of endometriosis terms and their appropriation by patients, from an applied perspective.

I will discuss challenges related to the necessary inclusion of heterogeneous texts to reflect the dynamics of term circulation in different speech communities, while keeping the corpus balanced and representative, and ensuring its comparability. The discussion will result in the proposition of a corpus composed of four sub-corpora that represent key stages of term circulation and that allow for the analysis of patients’ and laypeople’s understanding of terms through the comparison of the sub-corpora.

**References**

Delavigne, V., Picton, A., & Thibert, E. (2022). Socioterminologie et terminologie textuelle : L’expertise en questions. Proceedings of the 8th Congrès Mondial de Linguistique Française (CMLF 2022).

Gill, V. T., & Maynard, D. W. (2006). Explaining illness : Patients’ proposals and physicians’ responses. In D. W. Maynard & J. Heritage (Eds.), Communication in Medical Care : Interaction between Primary Care Physicians and Patients (p. 115 150). Cambridge University Press.

León-Araúz, P. (2015). Term Variation in the Psychiatric Domain : Transparency and Multidimensionality. In P. Hacken & R. Panocová (Eds.), Word Formation and Transparency in Medical English (p. 33 54). Cambridge Scholars Publishing.

**Keywords:** term circulation, endometriosis, comparable corpus, corpus compilation, representativeness

[6] ***Compiling and Annotating the First Electronic Learner Corpus of Romanian.***

Carmen Mîrzea Vasile (University of Bucharest & 'Iorgu Iordan - Al. Rosetti' Institute of Linguistics), Ana-Maria Barbu (University of Bucharest & 'Iorgu Iordan - Al. Rosetti' Institute of Linguistics), Enida Cincora Preda (University of Bucharest), Valentina Cojocaru (University of Bucharest & 'Iorgu Iordan - Al. Rosetti' Institute of Linguistics), Ramona Cătălina Corbeanu (University of Bucharest & 'Iorgu Iordan - Al. Rosetti' Institute of Linguistics), Mihaela Cristescu (University of Bucharest), Elena Irimia (University of Bucharest & Research Institute for Artificial Intelligence 'Mihai Drăgănescu'), Vasile Păiș (University of Bucharest & Research Institute for Artificial Intelligence 'Mihai Drăgănescu') and Monica Vasileanu (University of Bucharest & 'Iorgu Iordan - Alexandru Rosetti' Institute of Linguistics, Romanian Academy).

**Abstract.** Learner corpora have opened new horizons in foreign language learning research and practice. Romanian has not benefit yet from an electronic learner corpus; the LECOR project, running at the University of Bucharest, is aimed at filling this gap. The corpus, still in its early stages, will be an open access comprehensive resource, containing 4000 authentic texts produced by students in the intensive Romanian language course at the University of Bucharest, with various linguistic backgrounds, ranging from A1 to B2 level. 20% of the texts are transcripts of authentic audio recordings. Original samples will be available along the annotated texts. Students’ metadata refer to their demographic characteristics, language background and learning styles, and each sample will also have task metadata. Corpus annotation is currently in progress: all texts will be annotated for POS. The most challenging process is the manual annotation of 3500 sentences for morpho-syntactic errors, in the Relate platform.

**Keywords:** Romanian as a foreign language, corpus building, learner corpus, error annotation, learner corpus metadata

[7]***The Saudi Novel Corpus: Progress of compilation and preliminary results.***

Tareq Alfraidi (Islamic University of Madinah).

**Abstract.** There are several specialized Arabic corpora available that are built for various purposes, and they represent different genres. Nonetheless, online newspapers have been the dominant genre among the existing Arabic corpora (Al-Thubaity 2015, El-Khair, 2016). The Arabic novel genre, especially the Saudi novel, however, has been largely overlooked. This has led to the absence of Arabic corpus stylistics research. Such a lacuna has motivated us to create the first version of the Saudi Novel Corpus (SNCorpus) to contribute to the analysis of Saudi literary works and facilitate stylistic and linguistic studies in this area (Alfraidi et al., 2022).

In this poster, we aim to (i) present the procedures for expanding the current version of the corpus and (ii) report some results that emerged from the analysis of the content of the corpus. This is in order to demonstrate the value of using the corpus approach to exploring Arabic literature, especially Saudi novels.

To peruse the first aim, we will present the recent advancements achieved to improve the corpus quality. We will give particular attention to presenting the data collection procedure, text pre-processing, annotation, and interface programming.

The main recent additions to the corpus so far include:

1. Increasing the size from 3M words to 5M words.

2. Annotating the words of the corpus with POS tags and lemmatized them to expand the usability of the corpus.

3. Building a web interface that facilitates the search in the content of the corpus.

To fulfill the second aim, we will present the main results that emerged from the empirical investigation of the corpus. Using the functions available in the web interface, we generated a word list of the top 20 words and keywords and examined their concordance lines. As a result, we manage to spot several textual patterns.

Our hope is that this work will prove valuable to the Arabic stylistics and linguistics research communities and bridge the gap between the fields of corpus linguistics and Arabic literature.

**References**

Alansary, S.; Nagi, M.; Adly, N. Building an International Corpus of Arabic (ICA): Progress of compilation stage. In Proceedings of the 7th International Conference on Language Engineering, Cairo, Egypt, 5–6 December 2007; pp. 5–6.

Alfraidi, T.; Abdeen, M.A.R.; Yatimi, A.; Alluhaibi, R.; Al-Thubaity, A. The Saudi Novel Corpus: Design and Compilation. Appl. Sci. 2022, 12, 6648. <https://doi.org/10.3390/app12136648>

Al-Thubaity, A.O. A 700M+ Arabic corpus: KACST Arabic corpus design and construction. Lang. Resour. Eval. 2015, 49, 721–751.

El-Khair, I. A. (2016). 1.5 billion words Arabic corpus. arXiv preprint arXiv:1611.04033.

**Keywords:** Saudi novels, Corpus development, Arabic stylistics, Web interface

[8] ***Machine vs human transcription: issues in spoken learner corpus compilation.***

Mateus Miranda (University of Limerick).

**Abstract.** The architecture of a learner spoken corpus involves several steps, such as design, data recording, the creation of transcription parameters, and the validation process. Though speech has several elements that must be registered in the transposition to written form, making transcription by humans time consuming, software promises automatic transcription. This paper, which reports on a broader oral corpus compilation project of more than 100 recordings of learner language, seeks to discuss the transcription procedures established to reduce disagreements related to the guidelines. The methodology combines the analysis by automatic speech transcription software and human transcribers that included (a) transcription workshops and (b) transcription tasks. Finally, it presents the corpus transcription validation process that will be conducted to reduce disagreements related to the project criteria to prevent the lack of homogeneity in the final product.

**Keywords:** Learner corpus, Spoken corpus, Transcription

[9] ***Data Protection in Corpus-based Machine Translation.***

Xiaojun Zhang (Xi'an Jiaotong-Liverpool University).

**Abstract.** Machine translation is growing more and more powerful. It’s time that we stop to look at translation ethics and ask: is machine translation ethical? Before we answer that question, let’s look at Google Translate, a neural machine translation (NMT) tool. Users of the tool must agree to Google’s terms of service on its ownership of data. In other words, Google can use any data that you enter into Google Translate.

This is a fast way to translate paragraphs of text, emails, and even documents with machine translation, but it can be risky for companies dealing with confidential and secure information, especially when it comes to any type of legally sensitive material. We argue that there is an urgent need to develop ethical regulations to manage the data in MT activities, which will be helpful to protect users’ data in the digital age.

**Keywords:** data protection, machine translation, corpus-based

[10]***Discourse markers, a bastion of human translation against machine translation?***

Nathanaël Stilmant (Université de Mons).

**Abstract.** Despite tremendous progress, machine translation (MT) does not yet perfectly equal human translations (Maučec & Donaj, 2019). It still lacks creativity (Vanroy, 2021), for example in translating discourse markers (Meyer & Webber, 2013).

This exploratory study analysed a randomised sample of translations of the marker pair maar in Dutch and mais in French (100 occurrences per marker) from the Dutch Parallel Corpus (produced by professional human translators), and submitted it to the machine translation tool DeepL.

DeepL translated 97% of the mais and 97% of the maar literally, whereas human translators translated literally only 61% of the mais and 76% of the maar, opting in the remaining cases for different contrastive markers or for implicit relations. This suggests that the variety of translation strategies used to translate discourse markers remains an area where human translations are still distinct from MT.

**References**

Maučec, M. S., & Donaj, G. (2019). Machine Translation and the Evaluation of Its Quality. In A. Sadollah and T. S. Sinha (Eds.), Recent Trends in Computational Intelligence (pp. 143-163).

Meyer, T., & Webber, B. (2013). Implicitation of Discourse Connectives in (Machine) Translation. Proceedings of the Workshop on Discourse in Machine Translation (DiscoMT), 19-26.

Vanroy, B. (2021). Syntactic Difficulties in Translation (Doctoral dissertation, Ghent University).

**Keywords:** Machine Translation, Human Translation, Discourse Markers (Of Contrast), Dutch (maar), French (mais)

[11] ***Evaluating the Multi-Feature Tagger of English (MFTE): Challenges and implications for corpus tool evaluations and comparisons.***

Elen Le Foll (Osnabrück University) and Muhammad Shakir (University of Muenster).

**Abstract.** The reliability of multi-dimensional (MD; Biber 1988; 1995; Berber Sardinha & Biber 2014) framework largely depends on the linguistic features entered in such analyses. Not only do these need to be well chosen and operationalised, the software used to identify and count them also needs to be accurate.

We report on the formal evaluation of an open-source tagger for MD analyses: the Python version of the Multi-Feature Tagger of English (MFTE), which tags and counts 78 lexico-grammatical features, with an optional extended tagset of 64 additional features, including many semantic features. Our evaluation includes measures of recall and precision for every feature of the default tagset across a range of registers and varieties of English. These measures are compared to other popular MD taggers. Our results show that the MFTE performs well and offers a more transparent alternative that can easily be adapted to meet the needs of specific corpora and/or specialised registers.

We reflect on the limitations of the evaluation process, the need to establish best practices for evaluating taggers and their contribution to the replicability of MD studies.

**References**

Berber Sardinha, Tony & Douglas Biber (eds.). 2014. Multi-Dimensional Analysis, 25 Years on: A Tribute to Douglas Biber (Studies in Corpus Linguistics (SCL) 60). Amsterdam: John Benjamins.

Biber, Douglas. 1988. Variation across speech and writing. Cambridge: Cambridge University Press.

Biber, Douglas. 1995. Dimensions of Register Variation. Cambridge: Cambridge University Press.

**Keywords:** Tool evaluation, Automatic annotation, Inter-rater reliability, Multi-dimensional analysis (MDA), Register analysis, Reproducibility, Open Science

[12] ***Assembling EuReCo for contrastive research: The Polish piece.***

Piotr Banski (Leibniz-Institut für Deutsche Sprache, Mannheim), Nils Diewald (Leibniz-Institut für Deutsche Sprache), Marc Kupietz (Leibniz-Institut für Deutsche Sprache) and Beata Trawinski (IDS Mannheim).

**Abstract.** Together with several European partners, the Institute for the German Language (IDS) has been involved in an initiative known as the European Reference Corpus EuReCo (Kupietz et al., 2020). EuReCo is an emerging federated virtual corpus, with large user-definable comparable corpora across various languages and with an infrastructure supporting contrastive research. The core of the infrastructure is KorAP (Diewald et al., 2016), a scalable open-source platform supporting the analysis and visualisation of properties of texts annotated by multiple and potentially conflicting information layers, and supporting several corpus query languages.

So far, EuReCo comprises the German Reference Corpus DeReKo (Kupietz et al., 2018), the Reference Corpus of Contemporary Romanian Language (Barbu Mititelu et al., 2018) and the Hungarian National Corpus (Váradi, 2002).

The poster presents the result of a pilot project aiming at enriching EuReCo with the National Corpus of Polish (Przepiórkowski et al., 2010). We look at the process of integrating the Polish dataset into EuReCo and demonstrate some of the possibilities for expanding the scope of contrastive studies after the Polish dataset has been added to EuReCo.

**References**

Barbu Mititelu, V., Tufiş, D., Irimia, E., 2018. The Reference Corpus of the Contemporary Romanian Language (CoRoLa), in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan.

Diewald, N., Hanl, M., Margaretha, E., Bingel, J., Kupietz, M., Bański, P., Witt, A., 2016. KorAP Architecture ― Diving in the Deep Sea of Corpus Data, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16). European Language Resources Association (ELRA), Portorož, Slovenia, pp. 3586–3591.

Kupietz, M., Diewald, N., Trawiński, B., Cosma, R., Cristea, D., Tufiş, D., Váradi, T., Wöllstein, A., 2020. Recent developments in the European Reference Corpus EuReCo. Transl. Comp. Lang. Corpus-Based Insights Sel. Proc. Fifth Using Corpora Contrastive Transl. Stud. Conf. Louvain--Neuve Press. Univ. Louvain, Selected Proceedings of the Fifth Using Corpora in Contrastive and Translation Studies Conference 257–273.

Kupietz, M., Lüngen, H., Kamocki, P., Witt, A., 2018. The German Reference Corpus DeReKo: New Developments – New Opportunities, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan.

Przepiórkowski, A., Górski, R.L., Laziński, M., Pęzik, P., 2010. Recent Developments in the National Corpus of Polish, in: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10). European Language Resources Association (ELRA), Valletta, Malta.

Váradi, T., 2002. The Hungarian National Corpus, in: Proceedings of the Third International Conference on Language Resources and Evaluation (LREC’02). European Language Resources Association (ELRA), Las Palmas, Canary Islands - Spain.

**Keywords:** contrastive linguistics, comparable corpora, EuReCo

[13] ***A Small Corpus of Scene-Annotated Narrative Texts.***

Tarfah Alrashid (The University of Sheffield) and Robert Gaizauskas (The University of Sheffield). *ScANT:*

**Abstract.** We present the first scene annotated English dataset based on SceneML[1]. SceneML is a framework for annotating scenes in narrative text, where a scene is defined in sceneML as a “unit of narrative that has three basic elements: time, location, and main characters”. Any change in these elements indicates a change in the scene. Annotators were asked to annotate scene description segments (SDSs) in given texts; and scene transition segments (STs).

The dataset is composed of text from children’s stories and novels from Project Gutenberg. Selected chapters from Bunnies from the Future, The Wonderful Wizard of Oz, Pride and Prejudice, Sherlock Holmes, A Tale of Two Cities, and The Great Gatsby have been annotated for SDSs and STs by two annotators. The corpus consists of fourteen chapters and has 2796 sentences. Inter-annotator agreement scores have been computed for all the 14 chapters. The poster will include a fuller description of the annotation scheme and the annotated corpus, in addition to the preliminary results of an automatic scene segmentation system trained and tested on the corpus.

**Reference**

[1] Gaizauskas, Robert, and Tarfah Alrashid. "SceneML: A proposal for annotating scenes in narrative text." Proceedings of the 15th Joint ACL-ISO Workshop on Interoperable Semantic Annotation. 2019.

**Keywords:** Scene, Annotation, Corpus, Text, Narrative

[15] ***MultiTransRT: Multilingual Parallel Corpus and Platform for Translation Research and Teaching in China.***

Baorong Huang (Institute of Corpus Studies and Applications, Shanghai International Studies University), Xiaoqian Li (Institute of Corpus Studies and Applications, Shanghai International Studies University) and Kaibao Hu (Institute of Corpus Studies and Applications, Shanghai International Studies University).

**Abstract.** In the present period, where translation is playing an increasingly significant role in our society, translator teaching or training is crucial. It has been common practice to include corpora as part of training for the translation profession, and a wide variety of corpora such as parallel corpora, comparable corpora, and monolingual corpora, have been used to enrich translators’ awareness of subtleties across languages and enhance translators’ mastery of translation strategies. Based on years of translation teaching and training practice, we compiled a multilingual parallel corpus and developed a supporting corpus platform. The corpus contains Volume I, Volume II and Volume III of the Governance of China by President Xi Jinping in more than ten languages, annotated with translation strategies that can be applied immediately in translation instruction. The supporting platform is comprised of three functional modules: an aligner, a research module, and a teaching module containing parallel article library, knowledge base, and quotation bank. The aligner supports project-based parallel document alignment. The research module supports word frequency, Ngram, collocation, and related visualization. In the teaching module, lecturers can easily find relevant authentic examples by searching for a specific translation in parallel articles, examine the context of certain translation strategies, and summarize the translation strategies used in political translations. In addition, the knowledge base and quotation bank contain rich resources for understanding related Chinese idioms, quotations and concepts. With the multilingual corpus and the platform, researchers can do research on translation universals across several languages, on the strategies in the political translation, or comparative research with the bilingual or multilingual materials, complemented by detailed explanations of Chinese concepts.

**Keywords:** political translation, multilingual parallel corpus, corpus platform, translation research and teaching

[16] ***Validating Terminologies and Phraseological Units Retrieved from Specialized Comparable Corpora in Lexical Semantics: An Interactive Method.***

Hsin-Yi Lien (Ming Chuan University) and Hsieh-Chih Lai (National Academy for Educational Research).

**Abstract.** The methods of extracting terminologies and phraseological units from comparable corpora in previous studies were inclined to employ statistical machine translation or computational analysis. However, it was apparently insufficient for ensuring semantic level of obtained terms. Accordingly, the present study will apply an interactive method for cross validation of the quality of retrieved terms and phraseological units from specialized comparable corpora. The specialized comparable corpora consist of a Buddhist English Corpus and a Buddhist Chinese Corpus. The proposed interactive method includes filtering the terms and phraseological units with criteria; validating terms and phraseological units with different references sources including Google search engine, English dictionaries, Chinese dictionaries, and Pali dictionaries; implementing various statistical measures such as absolute frequency, log-likelihood (LL), odds ratio (OR) for ensuring the distinctness of obtained terms, and machine translation for comparing terminologies and phraseological units. Additionally, the distinct terminologies and phraseological units extracted from specialized Buddhist comparable corpora are examined in lexical semantics. Furthermore, the change of trend in Eastern and Western Buddhist literature will be explored through comparing the extracted terms and phraseological units from the Buddhist comparable corpora. Based on the results, the study will provide some suggestions for future studies.

**Keywords:** terminology, phraseological units, comparable corpora, lexical semantics

[17] ***A Distributional Comparison between FOLK and DeReKo.***

Marc Kupietz (Leibniz-Institut für Deutsche Sprache), Peter Fankhauser (Leibniz-Institut für Deutsche Sprache) and Josef Ruppenhofer (Leibniz-Institut für Deutsche Sprache).

**Abstract.** The goal of this study is to compare language use in spoken vs. written contemporary German. To this end, we contrast a orthographically normalized variant of version 2.18 of the Forschungs- und Lehrkorpus Gesprochenes Deutsch FOLK with 3.4M tokens (Schmidt 2017) with the DeReKo-2017-II edition German Reference Corpus with 33.0G tokens (Kupietz et al. 2018). We train a structured skipgram word2vec model (Ling et al. 2015) for DeReKo and retrain it on FOLK to arrive at comparable word embeddings for spoken language. On this basis we compare relative frequencies of words in DeReKo and FOLK, and analyse word embeddings for maximum displacement between their written and spoken variants together with their paradigmatic and syntagmatic neighbourhood. First results indicate that the top words w.r.t. difference in relative frequency (including particles and personal pronouns) are also maximally displaced and accordingly change their syntagmatic neighbourhood. However, their paradigmatic neighbourhood remains fairly stable, i.e., displacement occurs for whole paradigmatic clusters rather than on individual words.

**References**

Kupietz, Marc/Lüngen, Harald/Kamocki, Paweł/Witt, Andreas (2018): The German ReferenceCorpus DeReKo: New Developments – New Opportunities. In: Proceedings of the 11th InternationalConference on Language Resources and Evaluation (LREC 2018). Miyazaki/Paris:European Language Resources Association (ELRA), pp. 4353-4360.

Ling, Wang / Dyer, C. / Black, A. / Trancoso, I. (2015): Two/too simple adaptations of word2vec for syntax problems. In Proc. of NAACL.

Schmidt, Thomas (2017): Construction and Dissemination of a Corpus of Spoken Interaction –Tools and Workflows in the FOLK project. Journal for Language Technology and Computational Linguistics (JLCL 31/1), pp. 127–154.

**Keywords:** distributional semantics, word embeddings, spoken vs. written

[18] ***Analyzing French and Swedish web registers using text dispersion keywords.***

Saara Hellström (University of Turku).

**Abstract.** The internet features a wide variety of registers (Biber, 1988), some of which are exclusive to it, such as blogs or discussion forums, and some of which have their origins in the printed media, like news or reviews. English web registers have been studied in detail (Biber & Egbert, 2018), but online registers in other languages remain largely unexplored. This research will expand the study of web language use to French and Swedish web registers by describing them and analyzing their keywords using text dispersion (Egbert & Biber, 2019). The new data has been recently finished and consists of approximately 3,200 texts in both languages. The register annotations follow the same scheme as in the Corpus of Online Registers of English, CORE, (Biber & Egbert, 2018) and FinCORE (Laippala et al., 2019), which makes the corpora comparable and comparative research possible.

**References**

Biber, D. (1988). Variation Across Speech and Writing. Cambridge: Cambridge University Press.

Biber, D. & Egbert, J. (2018). Register Variation Online. Cambridge: Cambridge University Press.

Egbert, J. & Biber, D. (2019). Incorporating text dispersion into keyword analyses. Corpora 14, no. 1: 77–104.

Laippala, V., Kyllönen, R., Egbert, J., Biber, D., & Pyysalo, S. (2019). Toward multilingual identification of online registers. In M. Hartmann & B. Plank (Eds.), Proceedings of the 22nd Nordic Conference on Computational Linguistics (pp. 292– 297). Linköping University Electronic Press. <https://www.aclweb.org/anthology/W19-6130.pdf>

**Keywords:** web registers, text dispersion, French, Swedish

[19] ***A Spanish-English parallel corpus from financial reports.***

Yanco Torterolo (Universidad Autonoma Madrid), Sofia Roseti (Universidad Autonoma Madrid), Blanca Carbajo Coronado (Universidad Autónoma de Madrid) and Antonio Moreno-Sandoval (Universidad Autonoma Madrid).

**Abstract.** We present the process of compiling a parallel corpus from financial reports in Spanish and their translation into English —downloaded from the websites of the IBEX-35 companies. Our aim is to create a segmented, aligned bilingual corpus to carry out linguistic and translation studies and to create linguistic resources for AI.

The extraction and structuring of the information always pose the biggest challenges when compiling a corpus from PDF documents, as the information is presented in several columns with a non-linear organisation, which hinders the automatised extraction of the text.

We showcase our method for extracting the narrative elements, the subsequent cleaning of the text and the alignment of the paragraphs in Spanish and English. The result is a CSV file containing both languages. We used 15 bilingual reports resulting in 1,678,426 words in Spanish and 1,452,636 words in English, and 56,170 segments in Spanish and 56,813 segments in English.

**Keywords:** bilingual corpus, parallel corpus, compilation, financial domain

[20]***Design of a bilingual school textbook corpus specialized in science.***

Miriam Buendia-Castro (University of Granada).

**Abstract.** Bilingual programmes, especially English/Spanish ones, have grown rapidly in primary education in Spain. Subjects such as Science are taught in English in bilingual schools. One of the main problems faced by students and their families is the lack of knowledge of the foreign language and the lack of resources to support this new teaching-learning environment. This research outlines the methodology for the design of an English-Spanish school dictionary, specialised in science, aimed at supporting bilingual education in Spain. To this end, a comparable English-Spanish corpus has been compiled derived from the textbooks of one of the most popular publishers in Spain. For the English subcorpus, the textbooks available in bilingual schools were chosen and for the Spanish subcorpus, the textbooks available in non-bilingual schools. More specifically, the textbooks of the subject of Science for 3rd, 4th, 5th and 6th grade (KS2 in the UK) were used. It should be highlighted that, at the school level, to date, few school textbook corpora have been built (Candarli 2022). This paper describes the design of this bilingual corpus, its content and size, and how it was analysed.

**Reference**

Candarli, D. 2022. "Corpus data and method". In The linguistic challenge of the transition to secondary school (edited by A. Deignan, D. Candarli and F. Oxley). Routledge.

**Keywords:** sience, primary school, comparable corpus, school textbook corpus, science

[21] ***Good Listenership in EFL coursebooks: A context-driven Corpus Pragmatics Enquiry.***

Kevin Frank Gerigk (Lancaster University).

**Abstract.** Listening comprehension texts in EFL textbooks for adult and further education in Germany and Austria claim to expose the L2 learner of English to authentic, real-life dialogues. This study investigates how, in this context, the listener is represented in dyadic conversation - as a passive recipient of information or as an interactive participant in dialogue. Listening in English is not a passive undertaking in casual conversations, but a highly active endeavour that sees the listener use a range of linguistic devices to signal continued attention and have an important phatic function such as listener response tokens, sentence-completion routines and which-comment clauses (McCarthy & O’Keeffe, 2014). To investigate how good listenership is represented in listening comprehension compared to naturally occurring English, I devised a context-driven, functions-to-form annotation scheme based on corpus pragmatics research, predominantly suggested by O’Keeffe et al. (2007). My annotation scheme focusses on contextual units, in which good listenership is likely to occur, and allows me to annotate its features in dependence on how

the continuation or termination of the active turn is influenced. This approach has enabled me to a) capture a more inclusive picture of what constitutes good listenership and b) minimise the hazard of false positives. My annotated dataset allows me to answer the following two broad research questions: 1) Is there a statistically significant difference in the occurrence of features of good listenership between the listening comprehension texts and ‘real’ spoken English? And 2) Do features of good listenership occur in the same contextual units in the listening tasks as they do in the natural spoken data? By answering these questions, I aim to identify potential shortcomings in the teaching of listening and speaking and suggest ways for improving the representation of listenership in EFL coursebooks. In this talk, I will present my context-driven research approach and my initial findings.

**References**

McCarthy, M., & O’Keeffe, A. (2014). Spoken Grammar. In M. Celce-Murcia, D. M. Brinton, & M. A. Snow (Eds.), Teaching English as a Second or Foreign Language (4th ed., pp. 271–287). (Pre-published version).

O’Keeffe, A., McCarthy, M., & Carter, R. (2007). From corpus to classroom—Language use and language teaching. Cambridge University Press.

**Keywords:** Good Listenership, Corpus Pragmatics, ELT materials, L2 Listening Comprehension

[22] ***Insights into pedagogical applications of corpus-based biomedical wordlists in ESP settings.***

Neila Cheriet (University of Limerick).

**Abstract.** Corpus linguistics plays a significant role in specialized language teaching in academic disciplines through developing subject-specific wordlists. Coxhead’s Academic Word List has been used for informing general EAP coursebooks while the development of more specialized wordlists such as the Medical Academic Vocabulary List (MAVL) is dedicated to biomedical disciplines and has the potential to inform specialized vocabulary teaching. However, little research has been done about how teachers and learners can exploit wordlists in specialized pedagogical settings. This research focuses on the exploitation of the MAVL to support the teaching and learning of academic medical lexis. This poster presents a web-based resource which was developed to allow ESP teachers to find authentic texts containing MAVL words to develop learners’ knowledge of key medical lexis through pedagogically designed tasks. These include a word-finder, corpus-based exercises, gap-fill and quizzes. Later, ESP teachers will be surveyed to evaluate the effectiveness of the web-based resource.

**Keywords:** Specialized wordlists, Academic Word List, MAVL, Web-based resource, Word-finder, Corpus-based exercises

[23] ***Mapping shared lexical bundles onto rhetorical moves in research articles: A comparative study of paradigmatic variation.***

Mei Yang (University of Helsinki).

**Abstract.** Researchers have recently explored the influence of qualitative, quantitative, and mixed methods research paradigms on lexical bundles in research articles (RAs). However, these studies compared frequencies of broad function categories in terms of whole texts, lacking a comparison of the individual bundles under each function category in specific parts within texts. Besides, concordance analyses of lexical bundles’ discourse functions in these studies were primarily performed for distinctive bundles by listing several examples, lacking a detailed exploration of the specific contexts in which shared bundles are used in in-text parts in three paradigms. To address these issues, this study seeks to combine lexical bundles with rhetorical moves based on a 1,996,411-word corpus consisting of qualitative, quantitative, and mixed methods nursing RAs. This study compares the frequencies of shared bundles in not only whole texts but also each move where the shared bundles occur, and analyses how shared bundles are utilised to achieve communicative purposes in each move. Significant differences are found in the frequency of shared bundles in moves, and distinctive functions are found of shared bundles between mixed methods and qualitative nursing RAs, and between mixed methods and quantitative nursing RAs. This variation may be related to the epistemologies underlying the three research paradigms.

**Keywords:** lexical bundles, research articles, rhetorical moves

[24] ***Causality in Spanish financial discourse.***

Blanca Carbajo Coronado (Universidad Autónoma de Madrid).

**Abstract.** This work aims to analyse the frequency and distribution of linguistic strategies of causality in Spanish financial discourse, as well as the ratio of cause versus effect devices (see Sun-Young 2022) using the corpus manager tool Sketch Engine. For this purpose, I rely on a corpus of corporate annual reports in Spanish (Moreno et al., 2020), with more than 5 million tokens.

The starting point is the study of Gozalo (2004), who compiles all the forms of expression of cause and effect in Spanish. This work will focus specifically on nouns (efecto ‘effect’), verbs (provocar ‘trigger’), causative periphrases (hacer llorar ‘make sb cry’), prepositions (por ‘because of’), complex prepositions (a fuerza de ‘by dint of’) and conjunctions (pues ‘because’).

Ideally, this study will provide a comprehensive collection of patterns used in financial texts, forming the basis of a future annotation guide of causal relations.

**Keywords:** causality, cause, effect, finance, Spanish, corpus

[26] ***MexLeC. A spoken and longitudinal corpus of Mexican beginner to advanced learners of English.***

Ana Abigahil Flores Hernandez (Universidad Autonoma de Queretaro).

**Abstract.** The present study reports the process of designing and collection a spoken and longitudinal corpus of Mexican university learners as well as a brief description of the data obtained throughout two years of work as part of a Postdoctoral research project. This project aim is to collect a national data base to be used in the development of learner centred tools and materials for ELT and as an empirical base and complementary methodology in Mexican SLA research (Meunier, 2021; Guilquin, 2015).

To elicit spoken production, it has been selected monological and non-interactive tasks to capture learner interlanguage extended turns. The tasks are contained in a 10-16 minutes semi-guided interview which is expected to be applied every year, following learners acquisition process during 4-5 years. The rationale of these tasks are descriptors from CEFR (Council of Europe, 2018) matched with an analysis of tasks used in examinations and 140 currently available English learner corpora, and the text types in (Biber, 2004) as internal and external considerations on representativeness.

Interviews were videorecorded using the Zoom app and the transcription guidelines have been adapted from Gablasova, Brezina and McEnery (2019) and Granger et al (2009). By July 2023 the corpus will include cohorts from four different state universities holding proficiency levels from A1 to B2. the first one, with three years tracking time (3 samples); the second, with two years tracking time (2 samples) and; the third and fourth ones, with a first sample collected.

By now, the size of the corpus is up to 200,000 tokens and some of the most interesting preliminary findings are the (expected) low type/token ratio scores; the wide use of fillers and pauses followed by elaborated chunks; and, the dissimilar features produced in the narrative task of the ones expected to be distinctive of this text-type.

**References**

Biber, D. (2004) Conversation text types: A multi-dimensional analysis. 7es Journées internationales d’Analyse statistique des Données Textuelles

Council of Europe 2018. Common European Framework of Reference for Languages: Learning, Teaching, Assessment; Companion Volume with New Descriptors. Strasbourg, Language Policy Division: Cambridge University

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

Guilquin, G. (2015) From design to collection of learner corpora. In Granger, Sylviane, Gilquin,

Meuner, F. (2021) Introduction to learner corpus research. In Tracy-Ventura, N. and Paquot, M., eds. The Routledge Handbook of Language Acquisition and Corpora. London: Routledge.

**Keywords:** Learner Corpora, Spoken Corpora, Longitudinal Corpora

[27] ***A Study on the Information Types of Dictionary Definition Sentences depending on the Generic Concept Categories of a Terminology: With a Focus on the Corpus of Terminology Definitions for the Physics Field in the Standard Korean Dictionary.***

Jun Choi (Chonnam National University), Jinsan An (Kyungpook National University) and Minkyu Sung (Kyungpook National University).

**Abstract.** The present study focuses on analyzing the content category of corpus for terminologies and definition sentences, unlike previous studies on terminologies and definition sentences that focused mainly on clarifying the formal pattern of definition sentences. The purpose of the study is to classify in abstract the generic concept of individual scientific terminology and then to identify the connection between a fixed generic concept category and the information items of definition sentence description by analyzing the information types of dictionary definition sentences depending on the type of generic concept. To this end, the study systematically organized generic concepts for roughly 7,000 physics terminology listed in the Standard Korean Dictionary by critically reviewing previously constructed systems of semantic networks for Korean vocabulary and then building a dictionary definition corpus(on the scale of 200,000 words) for the terms to categorize information types that contain definition entries. The work of subclassifying information communicated by terminology definition sentences makes use of the results of previous studies on information types for terminology definition sentences by Flowerdew(1992) and Choi(2019). In this way, it is expected to be used as a useful standard model for constructing terminology dictionaries.

**References**

Choi, J. 2019. Information Types and Patterns Analysis of the Supplementary Definitions of Scientific Terms in the Standard Korean Language Dictionary Definition Corpus, KOLEX 34: 127-157.

Flowerdew, J. 1992. Definitions in Science Lectures, Applied Linguistics 13(2): 202-221.

**Keywords:** scientific terminology, information types, generic concept, dictionary definition corpus

[28] ***A study on the usage prediction of Chinese aspect particles in Narratives using corpus and deep learning language models.***

Injung Jung (Korea Military Academy).

**Abstract.** Mandarin Chinese, as a paratactic grammar, the usage patterns of aspect particles are treated as optional in sentence/discourse. They are used as multifunctional forms indicating temporal meaning and epistemic modal, which is considered a challenging problem in grammar teaching. This study aims to examine the usage patterns of Chinese aspect particles Le(perfective), Zhe(imperfective), and Guo(experience) in narrative discourse using corpus-based and Deep-Learning-based language models. For this study, we built the custom corpora consisting of 60 natural stories(280, 344 character sets) and media scripts from Media Language Corpus(http://ling.cuc.edu.cn/rawpub/).

Based on the domain knowledge of Chinese discourse grammar, this study used both corpus-based and deep-learning-based language models to analyze the usage of aspect particles in narratives which reveals the flow of time. In a deep-learning-based language model, we investigate how accurately the BERT model can predict Chinese aspect particles. We believe that this study is not only meaningful in the field of Natural Language Processing, but also provides insight into Chinese grammar research or language education. This methodology can provide insight into establishing the Chinese grammar prediction system, which can help Chinese learners improve their skills by showing them grammatically correct and proper expressions by context.

**Keywords:** Chinese aspect particles, usage prediction, discourse grammar, narratives, corpus-based analysis, deep-learning-based language models

[29] ***Premodification patterns of the noun phrase in learner English writing across cohorts.***

Pamela Andrea Saavedra Jeldres (The University of Warwick).

**Abstract.** This is a learner corpus study in progress and it analyses the grammatical complexity of the noun phrase, based on Biber, et al’s hypothesised developmental index (2011). This presentation focuses on features of premodification, namely attributive adjectives and nouns as premodifiers, and how these features are associated with the learners’ cohorts (year 3, 4, 5 undergraduate programme) in academic writing. Dataset were taken from the CELTEC, a 246,808-word learner corpus compiled with the writings of future teachers of English in Chile. Preliminary results show that there is an association between the use of the attributive adjectives and pre-modifying nouns and the cohort or year of study of the learner. A qualitative analysis of the data showed some repeated cases of phrasal modification, specific adjective-noun and noun-noun sequences and in regard to syntax, a tendency of the noun-noun sequences to be placed in object position in prepositional phrases.

**Keywords:** Noun phrase commplexity, Learner corpus research, Grammatical complexity development, learner writing

[30] ***A Corpus-based Study of the Use of Summative Connectors in Arab Leaners Academic Writing.***

Amani Alonayzan (The University of Leeds).

**Abstract.** The aim of the present study is to examine how Arab learners and English native speakers (British and English) use discourse connectors to link their sentences to achieve cohesion when writing Academics texts which is considered problematic for EFL writers. The focus is first on how frequently advanced Arab learners use a list of summative connectors following the taxonomy adopted from Quirk's et al., (1985). Second, how Arab learners’ use of these connectors differ to native speakers. It is a corpus based study, and there corpora are used; a sub-corpus of the first Arab Learner English Corpus (ALEC) consisting of Argumentative essays. The reference corpora are two sub corpora of the International Corpus of Learner English (ICLE); one is (USARG) which refers to the American segment and (BRSUR) refereeing to the British. Both quantitative and qualitative approaches were used for the analyses and a Log Likelihood (LL) statistical test was conducted to compare the frequencies. The findings reveal that Arab students generally overuse summative connectors in comparison to native speakers.

**Keywords:** Arab learners, corpus based study, discourse connectors

[31] ***Investigating formulaic sequences in L2 learners’ academic writing in the EMI programmes.***

Tanjun Liu (Xi'an Jiaotong Liverpool University).

**Abstract.** English-medium instruction (EMI) has been increasingly adopted in higher education institutions in China. One of the crucial factors to implement cogent EMI programmes is students' English proficiency. Formulaic sequences are a vital component of language competence. However, formulaic sequences remain a challenge to L2 learners at different proficiency levels. This paper, therefore, aims to investigate the use of formulaic sequences in EFL learners’ disciplinary writing in EMI settings. A corpus of learners’ writing was analysed using Sketch Engine. The writing was collected from 60 undergraduates majoring in Business, English and Science and Technology programmes over their four years of study at an EMI university in China. The preliminary results showed disciplinary variations in the use of formulaic sequences in the different programmes. There was also some development in using formulaic sequences over time. These findings contribute to our understanding of improving learners’ use of formulaic sequences in disciplinary academic writing.

**Keywords:** formulaic sequences, language teaching and learning, academic writing, English-medium instruction

[33] ***Intensification in Written L2 Italian in South Tyrol schools.***

Stefania Spina (Università per Stranieri di Perugia) and Andrea Abel (EURAC Bolzano).

**Abstract.** Intensification is the use of any linguistic device that scales a quality, by establishing different degrees of that given quality (Bolinger 1972). The investigation of the mechanisms involved in its acquisition and use is relevant to Learner Corpus Research: the possibility of grading a quality adds expressive richness to linguistic productions and is a step forward in the way learners acquire a more sophisticated awareness of L2s (Hendrikx et al. 2019; Lorenz 1999; Pérez-Paredes, & Díez-Bedmar 2012). However, intensification in L2s is still under-researched.

In this poster we present some preliminary findings of a project that seeks to fill this gap, by investigating the acquisition and use of intensification in young learners of Italian in the bilingual context of South Tyrol (Fiorentini & Sansò 2017), based on the Kolipsi (Glaznieks et al. in preparation), a longitudinal corpus of Italian and German L2 learner texts from upper secondary schools.

**References**

Bolinger, Dwight. Degree Words, Berlin, Boston: De Gruyter Mouton, 1972.

Fiorentini, I. & Sansò, A. (2017). Intensifiers between grammar and pragmatics. A lesson from a language contact situation. In M. Napoli & M. Ravetto (eds.). Exploring Intensification: Synchronic, diachronic and cross-linguistic perspectives, Amsterdam: Benjamins, pp. 173–192.

Glaznieks, A., Frey, J.-C., Nicolas, L., Abel, A. & Vettori, C. (in preparation): The Kolipsi Corpus Family. A collection of Italian and German L2 learner texts from secondary school pupils.

Hendrikx, I., Van Goethem, K. & Wulff, S. (2019). Intensifying constructions in French- speaking L2 learners of English and Dutch: cross-linguistic influence and exposure effects. International Journal of Learner Corpus Research, Vol. 5:1, pp. 63–103.

Lorenz, G. 1999. Adjective intensification - learners versus native speakers: A corpus study of argumentative writing. Amsterdam: Rodopi.

Pérez-Paredes, P. & Díez-Bedmar M.B. (2012). The Use of Intensifying Adverbs in Learner Writing. In Yukio Tono, Yuji Kawaguchi and Makoto Minegishi (eds), Developmental and Crosslinguistic Perspectives in Learner Corpus Research, Amsterdam: Benjamins, pp. 105–124.

**Keywords:** Intensification, Learner Corpus Research, longitudinal corpus, L2 Italian

[34] ***Metaphorical variation in the language of learners of English as L2: A learner corpus based study.***

Marcos de Oliveira

**Abstract.** This research aims at investigating the correlation between the metaphorical production in the 25 subcorpora of the International Corpus of Learner English and the variables related to the populations represented by the corpus (e.g. country of origin and L1) and the tasks conducted during the text collection which, according to Granger et al (2020), may impact the learner’s language production. To this end, the subcorpora will be tagged for POS and analyzed under the perspective of the Conceptual Metaphor Theory (Lakoff & Johnson, 1980) and Corpus Linguistics approach. Next, the most frequent nouns, verbs, and adjectives along each subcorpus will be identified by a computer script so that their metaphorical meanings can be found and displayed in their respective concordance lines. Finally, the statistical correlation between the conceptual metaphors manifested in the learner´s linguistic metaphors and the variables related to the learners and text collection procedures will be verified.

**Keywords:** Conceptual Metaphor, English as L2, Metaphorical Variation, English teaching

[35] ***Linguistic Complexity as a Dimension of Foreign Language Proficiency: Written English Performance of Czech Secondary Technical School Students.***

Tomáš Lorman (Charles University - Faculty of Education, Department of English Language and Literature).

**Abstract.** The project focuses on complexity (rather than accuracy) of written production as a measure of L2 proficiency in English. The theoretical part explores how complexity may be assessed while the practical part aims to investigate how complexity in L2 writing may be improved.

The practical part employs the methodology of learner corpus research based on contrastive interlanguage analysis (Granger, 1998a) of works written by students at a Czech secondary technical school and A-level essays included in the Louvain Corpus of Native English Essays ‘LOCNESS’ (Granger, 1998b). This analysis is further supported by application of the New General Service List produced by Lancaster University (Brezina and Gablasova, 2015). The preliminary results show low level of structural complexity, limited vocabulary range and overuse of frequent words.

The results of the analysis are further used as a basis for systematic pedagogical interventions in a longitudinal experimental action research.

**References**

Brezina, V. & Gablasova, D. (2015). Is There a Core General Vocabulary? Introducing the New General Service List. Applied Linguistics, Volume 36, Issue 1, 1-22.

Granger, S. (1998a). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A. P. Cowie (Ed.) Phraseology. Theory, analysis, and applications (p. 145–160). OUP.

Granger, S. (1998b). The computer learner corpus: A versatile new source of data for SLA research. In Granger, S. (Ed.) Learner English on Computer (p. 3-18). Addison Wesley Longman.

**Keywords:** corpus linguistics, applied linguistics, complexity, written production, L2 proficiency, secondary technical school, learner corpus research, contrastive interlanguage analysis, pedagogical interventions, action research, experiment

[36] ***Student Writing and Assessment Practices: combining corpus linguistics and in-depth interviews to investigate written assessment in a distance university context.***

Maria Leedham (The Open University), Prithvi Shrestha (The Open University), Jackie Tuck (The Open University) and Dana Therova (The Open University).

**Abstract.** Distance learning has become increasingly prevalent, particularly post-Covid 19, yet little is known about distance students’ assessed writing. This pilot study aims to compile an innovative 4-million-word corpus of contemporary distance university undergraduate writing across four contrasting disciplines (Business, Engineering, History, Childhood Studies). Contextual student data such as disability, carer status, prior education and socio-economic status is also being collected and built into the corpus as text headers. Alongside the student writing corpus, a marker feedback corpus (comprising both on-script and summary comments) is being created. Two tutor-markers per discipline will be interviewed with a focus on specific student texts, informed by early findings from both corpora. This presentation will focus on the methodology used to build the two corpora and on analysis of early corpus findings combined with marker interviews. This project is also contributing to research into the implications for marking of Open AI’s GPT-3.

**Keywords:** student writing, CADS, interviews, assessment, distance learning, disciplines

[37] ***Playing Teacher: using corpus resources to design an intervention to support the mentoring of Early Career Teachers.***

Kathryn Spicksley (Oxford Brookes University).

**Abstract.** Teacher attrition is a significant and persistent problem in England, but research has shown that successful mentoring can increase commitment. In the postdoctoral project presented in this poster, I will use corpus resources to develop a gamified intervention to support and enhance the mentoring of Early Career Teachers (ECTs), thus demonstrating an innovative application of corpus linguistics within the field of Education.

The project involves a corpus-assisted discourse analysis to identify salient identity positionings of teachers in cultural and political discourse. Corpora include the BNC2014 and the TV Corpus, alongside a corpus of speeches by government ministers previously built by the researcher (the ‘DFEMS corpus’). Findings will be synthesised and used to develop a ‘top trumps’ style card game, which will support ECTs and their mentors to discuss teacher professional identity in a supportive discursive environment.

The poster will disseminate early findings and discuss possibilities for the intervention design.

**Keywords:** Corpus-assisted discourse analysis, Education, Professional identity, Applied linguistics

[38] ***Why Do We Use Disfluency Markers in Messaging? A Corpus-based Study on Discourse Markers in Online Messaging.***

Mikyung Baek (Kyungpook National University), Yelin Go (Kyungpook National University) and Kilim Nam (Kyungpook National University).

**Abstract.** In light of a recent study showing that the representative disfluency markers "uh" and "um" in English are rather more representative of ‘fluency markers,’ it seems appropriate to point out that "fluency and disfluency are two sides of the same coin” (Crible et al., 2017:71). Korean discourse markers frequently used in spoken language like A, Eo, and Mwe have also been classified as a typical types of "stuttering" in traditional grammar. However, if the speaker intentionally uses these in written instant message communication, then they can be interpreted as representing "ideal delivery” (Clark, 2006:245) rather than "stuttering.“ In this study, we first extract all types of discourse markers such as A, Um, Mwe, etc. appearing in a 0.4 million word Korean messaging corpus and analyze the distribution and usage patterns of each individual discourse marker. Second, by contrasting the results with the disfluency phenomena observed in spoken language, we discover messaging-oriented characteristics of discourse markers. In doing so, we are able to discuss the primary roles of disfluency and the functions of individual discourse markers in messaging communications. The results of this study are meaningful in that understanding text message disfluency can help reveal why speakers use discourse markers as a prominent feature in their written internet language, even though such usage goes against the principle of economy.

**References**

Crible, L., Degand, L., & Gilquin, G. (2017). The clustering of discourse markers and filled pauses: A corpus-based French-English study of (dis) fluency. Languages in Contrast, 17(1), 69-95.

Degand, L., Gilquin, G., Meurant, L., & Simon, A. C. (2019). Should ‘uh’and ‘um’be categorized as markers of disfluency? The use of fillers in a challenging conversational context. Fluency and disfluency across languages and language varieties, 4, 67.

Tottie, G. (2014). On the use of uh and um in American English. Functions of Language, 21(1), 6-29.

**Keywords:** disfluency, discourse marker, online messaging

[39] ***A corpus approach to discourses of learning in online communities.***

Viola Wiegand (University of Stirling).

**Abstract.** Online learning communities have become an important part of a wide range of educational contexts, from traditional educational institutions such as schools and universities to private providers of online skills courses (in such diverse areas as cooking, crafting, financial planning, programming, etc.). This poster presents the early stages of a project that applies a corpus approach to education by studying discourses of learning across such formal and informal online learning communities. A pilot study compares the representation of ‘learning’ in descriptions of online courses offered by “formal” educational providers (universities) and more “informal” private providers. The wider project aims to study discourses across the online course sector to examine conceptualisations of learning and patterns of “effective” learning interactions in different online learning communities. The project draws on approaches and insights from corpus-assisted discourse analysis and education, as well as existing corpus work on online communities (e.g. Collins, 2019; Mackenzie, 2020).

**References**

Collins, L. C. (2019). Corpus Linguistics for Online Communication: A Guide for Research. Routledge.

Mackenzie, J. (2020). Digital interaction. In S. Adolphs & D. Knight (Eds.), The Routledge Handbook of English Language and Digital Humanities (pp. 49–65). Routledge.

**Keywords:** corpus-assisted discourse analysis, applied corpus linguistics, online communities, digital interaction, online learning

[41] ***A Study of the Communist Party of China (CPC)'s Democratic Discourse (1937-1949).***

Jing Wang (Lancaster University) and Nianchun Li (Southwest Jiaotong University).

**Abstract.** The Communist Party of China (CPC) has formed a unique democratic discourse in the practice of promoting democracy.1937-1949 was a crucial period for the CPC to master the right of democratic discourse through discourse games with other parties. The democratic discourse in this period also gradually matured.In this paper, the important documents of the CPC from 1937 to 1949 are analyzed by Nvivo12plus.It is found that the three core democratic discourses of “New Democracy”, “the People's Democratic Dictatorship” and “Democratic Centralism” constitute the Communist Party of China (CPC)'s democratic discourse system from the aspects of society, state and political party.These three core democratic discourses are produced in different historical and social situations, and have their own discourse structures. Through different discourse propaganda, they have played different discourse effects.

**Keywords:** CPC, Democratic Discourse, New Democracy, Democratic Centralism

[42] ***Modelling political ideologies from parliamentary speeches.***

Otto Tarkka (University of Turku), Kimmo Elo (University of Turku), Filip Ginter (University of Turku) and Veronika Laippala (University of Turku).

**Abstract.** Ideological speech often expresses underlying attitudes and assumptions, and can affect all levels of discourse (Van Dijk, 2013). In our study, we explore how political ideology influences and, on the other hand, is linguistically represented in political speech. To this end, we use deep learning, a framework that allows us to examine the distinctiveness of ideologies and identify their linguistic realizations in our corpus (Kyröläinen & Laippala, 2023). We train a deep learning model with political speeches from 8 parties held in the Finnish parliament between 2000 and 2021 to predict the party affiliation of the speaker and then apply the model explainability method SHAP (Lundberg & Lee, 2017) to analyse the linguistic cues - keywords - that the model most relies on. Preliminary results show that the model discriminates between parties accurately (F1=0.679) and the linguistic cues provide insight into party ideologies.

**References**

Kyröläinen, A., & Laippala, V. (2023). Predictive keywords: Using machine learning to explain document characteristics. Frontiers in Artificial Intelligence, 5. <https://doi.org/10.3389/frai.2022.975729>.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. 31st Conference on Neural Information Processing Systems (NIPS 2017).

Van Dijk, T. A. (2013). Ideology and discourse. In M. Freeden., L. T. Sargent, & M. Stears (eds.), The Oxford handbook of political ideologies. Oxford: Oxford University Press.

**Keywords:** deep learning, ideology, political speech

[43] ***No straight talk here: A multi-level analysis of hedging strategies employed by the Fed Chair in press conferences.***

Zhipu Yang (Lancaster University).

**Abstract.** This study aims to analyze hedging strategies in a financial communication context. The American Federal Reserve System (“the Fed”) is frequently investigated by economists, but has rarely been examined in linguistics. Fed communication has a significant influence on market reactions and the sequence of economic events, and recent studies also find evidence that Fed communication and the anticipated alterations in the monetary rates affect overall lending rates, firm investments as well as stock returns. Therefore, studying the mechanisms of Fed communication is valuable. Understanding Fed communication helps reveal the way the Fed manipulates public perception and assists investors to make wise decision. From the perspective of interpersonal pragmatics, this analysis of Fed Committee press conferences corpus identifies the hedging strategies employed by the Fed Chair at the lexical, syntactic and discursive levels during Q&A sessions to elucidate how the Chair uses hedging as vehicle to manipulate the public's perceptions of the institution. The findings reveal the Chair uses words of uncertainty, formulations of abstract information, and indirect responses to avoid inconvenient questions and promote a positive image, which deviates from the traditional view of hedging as a mere politeness phenomenon and echoes growing interest in relational work and identity construction within interpersonal pragmatics. This study enriches our understanding of the mechanisms of Fed communication, broadens data sets of hedging within financial communication contexts, and offers novel insights into hedging rationales and frameworks as well as the function of hedging. Furthermore, it raises valuable questions about the trade-offs between increasing transparency and maintaining the institution's positive image. Finally, it provides both linguists and economists with a robust descriptive basis for critically assessing Fed communication.

**Keywords:** Fed communication, Hedging strategies, Words of uncertainty, Abstract information, Indirect responses, Interpersonal pragmatics

[44] ***Twitter COVID-19 infodemic narratives about the Brazilian Federal Supreme Court: A lexical multidimensional analysis of the disinformation discourse.***

Luciana Nogueirol Lobo Marcondes (Pontifícia Universidade Católica de São Paulo), Deise Prina Dutra (Universidade Federal de Minas Gerais), Katherine Oliva Ortolani (Pontifícia Universidade Católica de São Paulo), Rafael Fonseca de Araújo (Pontifícia Universidade Católica de São Paulo) and Tatiana Schimdt de Almeida Lopes (Pontifícia Universidade Católica de São Paulo).

**Abstract.** Misinformation is false or inaccurate information deliberately intended to mislead (World Health Organization, 2022). This study presents an analysis of the discourses constructed in Twitter posts to propagate disinformation around the performance of the Brazilian judiciary, through the Federal Supreme Court (STF), an important social and political actor, which handed down several decisions related to COVID-19 that affected the population health and governance. In order to understand this disinformation discourse to create tools that could help the population to detect this phenomenon, we compiled a corpus with more than 288k tweets in Portuguese from the hashtags #covid-19, #stf, #supremo, #ministros, among others. Then, a Lexical Multidimensional Analysis (Berber Sardinha, 2019, 2020; Berber Sardinha & Fitzsimmons-Dooley, in preparation, Clarke et. al. 2022) was employed and allowed the identification of lexical parameters of variation between texts, statistically detected through of factor analysis and interpreted in terms of discursive dimensions.

**Keywords:** Corpus Linguistics, Federal Supreme Court, Infodemic in COVID-19, Lexical Multidimensional Analysis

[45] ***A Corpus Assisted Multimodal Discourse Analysis of Public Signs in China During the Zero Covid-19 Policy: Monolingual, Visual narrative and Discursive Convergence.***

Yee Chin Gan (Universiti Malaya) and Minjun Park (Duksung Women's University).

**Abstract.** China has maintained a strict "Zero Covid-19" approach as the rest of the world enters the end of the pandemic in 2021-22. Tens of thousands of covid-19 testing booths have been set up across China to provide mandatory nucleic-acid tests for the contagious Omicron variant of coronavirus to Chinese citizens and residents. The purpose of this study is to conduct a Corpus Assisted Multimodal Discourse Analysis (CAMDA) on the public signs used at these test kiosks.

About 1,000 digital images were collected from the online social media, Weibo. The content in the images will be encoded and transcribed for machines-readable and corpus techniques analysis purposes. We will examine the recurring patterns in both written messages and visual encodings to understand the semantic integration of semiotic choices in multimodal phenomena that shape the symbolic ideological meanings associated with collective healthcare discourse in the ‘Zero Covid-19’ era.

**Keywords:** Corpus-assisted, Multimodality, Public signs, Zero Covid-19, Healthcare discourse analysis

[46] ***“Mom is not invincible”: A corpus-based critical discourse analysis of regretting motherhood posts on Chinese Weibo.***

Shuning Liu (Xi'an Jiaotong University).

**Abstract.** Feelings of regret affect mental health. Previous studies have analyzed parental regret on western online platforms. However, little study has explored regretting motherhood in China. This study aims to examine discourses of maternal regret in China. The study collected 1153 posts containing the hashtag “#Do you ever regret having been a mother#” (当了妈妈你后悔了吗) on Weibo, a Chinese popular social media platform. Using 444 posts denying maternal regret as the reference corpus, this study adopted corpus-based critical discourse analysis approach to investigate 330 posts implying or indicating regret. Preliminarily, four discourses expressing maternal regret were identified: the Cause discourse, the Action discourse, the Imagination discourse and the Right discourse. The findings can facilitate an understanding of Chinese mothers’ hidden regretting motherhood experience and their attempts to challenge overarching discourses of normative motherhood. The results can provide implications for healthcare professionals.

**Keywords:** regretting motherhood, normative motherhood, corpus-based critical discourse analysis, Chinese mothers, Weibo

[47] ***Environmental Image Construction in Environmentally Sensitive Companies’ Social Responsibility Reports.***

Song Guo (Lancaster University).

**Abstract.** With the development of environmental movements, the idea of sustainable development has taken root. In this context, environmentally sensitive companies are under enormous public pressure due to their eco-unfriendliness. To cope with the legitimacy crisis, more and more companies begin to issue corporate social responsibility (CSR) reports, informing stakeholders of their social and environmental impacts. Adopting a corpus-based critical discourse analysis, this study explores how Chinese environmentally sensitive companies represent environment and employ legitimation strategies in CSR reports, so as to construct an environmental image conforming to their interests. Beginning with the analysis of the collocates of “environment”, the study discusses how positive as well as negative topics are presented in the reports by examining the concordances and finally sums up the legitimations strategies. The study finds that discursive strategies are widely used in corporate environmental image construction. Thus, these high-risk companies are no longer polluters, but guardians of the environment.

**Keywords:** environmental image, corporate social responsibility report, discursive strategy

[48] ***Taboo Language and Incest in the UK press (2017-2022) – Finding Absence in Corpus Linguistics.***

Sophie Eyssette (University La Sapienza, Rome, in cotutelle with the University of Silesia, Poland) and Gavin Brookes (Lancaster University).

**Abstract.** Taboos have been prevalent in all societies throughout history. Taboos prescribe behaviours related to death, food, and sexuality, and one of the primary sexual taboos is incest. Incest is worth exploring from a sociolinguistic perspective as it is both a social and linguistic taboo.

Additionally, taboo language is often defined as swear words (Allan & Burridge, 2007; Pedraza, 2018; Bednarek, 2019). However, this paper aims to identify taboo words that are not swear words, but rather words considered too taboo to be spoken (Khairullina et al., 2020).

Therefore, this research aims to analyse the unspeakable. To do this, this study addresses the methodological challenge of finding absence in a corpus. The issue of finding absence has been discussed in corpus linguistics, and methods have been developed to compare corpora to see if certain elements are missing from one dataset to another (Partington, 2014; Schröter & Taylor, 2018).

However, in this study, absence is the main selection criterion for building a corpus on incest taboo. The corpus consists of eight British newspapers from October 13, 2017 to October 14, 2022, covering the period from the emergence of the #MeToo movement up to the first collection day. The goal is to find articles that discuss incest without using the word ‘incest.’ The methodological challenge in carrying out this corpus linguistics study is to find a missing word through a search query that uses specific search terms.

To this end, I will explain why the search terms are the broad terms ‘abuse’ and ‘father.’ I will also introduce an iterative approach to narrow down a corpus of 23,015 articles, comprising 28,187,466 words, to a corpus of 258 articles, comprising 3,827,015 words. I will discuss limitations and solutions, and present results comparing a corpus containing the word ‘incest\*’ to a corpus deliberately omitting it.

**Keywords:** taboo language, corpus linguistics, corpus design, media discourse, incest taboo

[49] ***How British corpora reflect gender-relation through forms of address.***

Michael Pace-Sigge (University of Eastern Finland).

**Abstract.** This chapter investigates the use of terms of address in relation to a number of female and male names. As research as early as the 1990s (see, amongst others, Acker, 1990; Lakoff and Lakoff, 1990; Tannen, 1994; Wodak, 1996; Schmid, 2015) has shown, there is a clear link in the discourse between work roles and gender. This corpus-assisted research will use both the BNC 1994 and BNC 2014 to provide a basis for a qualitative and diachronic look at the uses of such terms like professor, director, minister etc. The use of these corpora provides an empirical snapshot of the choice of address employed in Britain in the 1980s/1990s and then in the early 2000s – in particular in the newsprint of the time. It also gives insight in the concrete nesting (cf. Hoey, 2005) of female names as compared to male names in mainstream discourse; this enables the construction of how the readership is psychologically primed to connect positions of responsibility and learning with the idea of ‘maleness’.

Overall, this research shows that forms of direct address is comparatively rare in spoken discourse; in written discourse, the majority is found in news items. Crucially, while there is a notable shift reflecting greater participation of females in position of power. The shift away from all-male politics in the 1980s has given way to a more equitable outlook. 30 years later, a shift within academia is also noticeable. However, in the field of business, relatively little seems to have changed. In fact, the use of boss is used more than before as a referrer to males.

**References**

Acker, J. (1990). Hierarchies, jobs, bodies: A theory of gendered organizations. Gender & Society, 4(2), 139-158.

Brezina, V., Hawtin, A. and McEnery, T., (2021). The Written British National Corpus 2014–design and comparability. Text & Talk, 41(5-6), pp.595-615.

British National Corpus (BNC). (1994). Reference Guide. <http://www.natcorp.ox.ac.uk/docs/URG/> (last accessed 21/10/2022)

British National Corpus 2014: User Manual and Reference Guide, Version 1.1. (BNC2014). Available from <http://corpora.lancs.ac.uk/bnc2014/doc/BNC2014manual.pdf> (last accessed 21/10/2022).

Hoey, M. (2005). Lexical Priming. A new theory of words and Language. London: Routledge

Lakoff, R. (2003). Language, gender, and politics: Putting “women” and “power” in the same sentence. The handbook of language and gender, 161, 78.

Lakoff, R. T., & Lakoff, R. (1990). Talking Power. Basic Books.

Schmid, H. J. (2015). Does Gender-related Variation Still have an Effect, Even When Topic and (Almost) Everything Else is Controlled. Change of Paradigms– New Paradoxes: Recontextualizing Language and Linguistics, 31, 327.

Tannen, D. (1994). Talking from 9 to 5: How women's and men's conversational styles affect who gets heard, who gets credit, and what gets done at work. New York: William Morrow and Company.

Wodak, Ruth (1996) Power, Discourse, and Styles of Female Leadership in School Committee Meetings. In: Discourse and Power in Educational Organizations. Cresskill, N. J., Hampton Press, pp. 31-54.

**Keywords:** BNC family of corpora, diachronic change, gender-related variation, lexical priming, naming conventions