Aston University
BIRMINGHAM UK

BIRMINGHAM CITY University

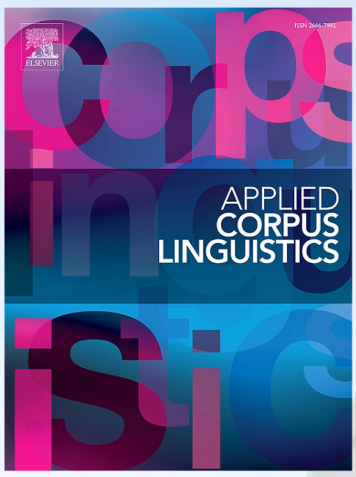
UNIVERSITY OF BIRMINGHAM

Corpus Linguistics 2025

The Thirteenth International Corpus Linguistics Conference

Birmingham, UK
30th June - 3rd July 2025

Book of Abstracts



BLOOMSBURY ACADEMIC

CAMBRIDGE UNIVERSITY PRESS

CLARIN
Common Language Resources and Technology Infrastructure

JOHN BENJAMINS PUBLISHING COMPANY

senso.

SKETCH ENGINE

CONTENTS

PLENARY PRESENTATIONS 3
GAVIN BROOKES 3
ELIZABETH HANKS 4
CHARLOTTE TAYLOR & ANNA MARCHI 5
LAURENCE ANTHONY 6
PASCUAL PÉREZ-PAREDES 7
PRE-CONFERENCE WORKSHOPS 8
THEMATIC PANELS 19
MONDAY 30th JUNE 19
TUESDAY 1st JULY 23
WEDNESDAY 2nd JULY 26
THURSDAY 3rd JULY 34
PAPER PRESENTATIONS 36
A 36
B 43
C 62
D 80
E 88
F 93
G 98
H 109
I 123
J 125
K 128
L 136
M 151
N 175
O 177
P 182
R 187
S 197
T 217
V 222
W 225
X 242
Y 244
Z 252
PRE-RECORDED PRESENTATIONS 262
POSTER PRESENTATIONS 265

PLENARY PRESENTATIONS

GAVIN BROOKES

Lancaster University

Putting corpora in context

In corpus linguistics, we have long grappled with the notion of context. As a nebulous concept, context has become highly polysemous, varying greatly across the range of subfields of applied linguistics that now make use of corpus approaches. In light of this potential variability, our quest to define context has given rise to a series of seemingly perennial questions that many working in the field are still grappling with to this day. These include questions such as, ‘How is context defined in corpus linguistics?’, ‘To what extent might corpora be considered devoid of context?’, ‘How do we integrate context into our corpora and into our corpus linguistic analyses?’, and ‘How much context do we need to ensure a rigorous and situated analysis?’.

In this plenary, I will address these and other questions like them as I take a birds-eye view of the debates surrounding, and the various perspectives on, the issue of context in corpus linguistics. I will consider some of the different ways in which context is conceptualised in corpus linguistics, as well as how these various conceptualisations present different challenges and, as such, give rise to distinct ways of embedding context into corpus analyses (and indeed, into corpora themselves). On this basis, I will reflect critically on long-standing claims about the ineffectiveness of corpus methods for interrogating context. In doing so, I will highlight some of the factors that have underpinned such claims and, looking to the future, I will consider how such factors might be attenuated through the ways we talk about and report on our data and analytical approaches.

ELIZABETH HANKS

Northern Arizona University

Sampling American English conversation: An exploration of corpus representativeness

Corpora are only valuable insofar as they represent the domain a linguist intends to study (Egbert et al., 2022). Designing highly representative corpora is particularly challenging for conversation, in part because the domain of conversation is difficult to operationalize. Previous studies have investigated incredibly heterogenous corpora that are all labeled 'conversation'; for example, unplanned/unedited spoken language, scripted interactions, filmed interactions, phone calls with strangers, and text messages. So, how much do we as a field know about what conversation really entails? And how much do we know about what corpora of conversation really represent?

To begin answering these questions, I present a comprehensive description of the register of conversation. This description draws on prototype theory to identify prototypical conversation texts and provide quantitative information about the situational and linguistic features that are most central to conversation (Hanks, 2025). I then introduce methods to quantitatively evaluate representativeness in terms of the domain considerations of a corpus. This evaluation compares continuous ratings for the contextual features of texts within a corpus to the central situational features of the target register.

I illustrate these ideas using The Lancaster-Northern Arizona Corpus of American Spoken English (LANA-CASE)—a new 10-million-word corpus of American English conversation that will be released open access in 2026 (Hanks et al., 2024). I evaluate several novel recruitment methods for their effectiveness at sampling situationally varied conversational language from diverse participants. I examine the extent to which each sampling method, as well as LANA-CASE as a whole, represents situational variation within the register of conversation (in terms of, e.g., planning, setting, communicative purpose) and demographic diversity in the U.S. (across ages, geographic regions, genders, and race/ethnicities). With appropriate adaptation, I show how the methods introduced in this talk can help us to better design, evaluate, and analyze corpora from any domain.

List of references

Egbert, J., Biber, D., & Gray, B. (2022). Designing and evaluating language corpora: A practical framework for corpus representativeness. Cambridge University Press.

Hanks, E., McEnery, T., Egbert, J., Larsson, T., Biber, D., Reppen, R., Baker, P., Brezina, V., Brookes, G., Clarke, I., & Bottini, R. (2024). Building LANA-CASE, a spoken corpus of American English conversation: Challenges and innovations in corpus compilation. Research in Corpus Linguistics, 12(2), 24-44. https://doi.org/10.32714/ricl.12.02.03

Hanks, E. (2025). Mapping out American English conversation: Central and peripheral features of intra-register variation [Doctoral dissertation, Northern Arizona University]. ProQuest Dissertations Publishing.

CHARLOTTE TAYLOR & ANNA MARCHI

[1]University of Sussex; [2]University of Bologna

What counts?

Counting has always been at the heart of corpus linguistics and is what unites us as a community. The papers at CL2025 range across many areas in linguistics but what we have in common is that we will all be counting something and somehow. The premises of corpus linguistics being that what counts is how language is used - and that counting how language is used reveals aspects of language that are invisible to the 'naked eye'. This invisibility may come about because a pattern is either so large, so diffuse, or so small that human perception alone cannot measure it. The centrality of counting is so ingrained that we rarely step back to look directly at our 'numerical habits' – and in the well-established tradition of reflexivity in CL - that is exactly what we would like to take the opportunity do in this plenary. We open up the space for reflection by asking five questions about what counts, asking: Why do we count? What do we count? How do we count? What counts as a big number? What shouldn't we count? Addressing these questions shows that counting is a theory-laden process because of all the decision making which underpins it and, at the same time, unveils its creative power.

LAURENCE ANTHONY

Waseda University

Stochastic parrots meet corpus linguists: Understanding language in the age of AI

The recent revolution in AI has led to astonishing innovations in language processing and generation, as well as the rapid advancement of multimodal human-human, human-machine, and machine-machine interfaces. Large Language Models (LLMs) have now become ubiquitous, with seemingly every device and Internet service integrating them in some way. However, the AI revolution has also raised very important questions about where, when, and how these models can be used safely, reliably, and responsibly. More fundamentally, many of the computer scientists who build these models still struggle to understand their exact nature. Are LLMs simply stochastic parrots that regurgitate language from vast Internet corpora, or are they more profound in nature, perhaps capturing some of the essence that makes human language so special?

In the talk, I will explore these questions by building on insights gained from over half a century of corpus linguistics research combined with knowledge of machine learning, natural language processing, and recent AI innovations. First, I will outline the basic architecture of LLMs, including the latest reasoning models such as DeepSeek-R1, which helps us to understand their strengths, weaknesses, and innate biases. Next, I will reflect on how LLMs align with models of human language processing, drawing on insights from conversation analysis, pragmatics, discourse analysis, and other areas of applied linguistics. These insights can help us to understand the potential for LLMs to uncover and explain patterns in human language within and across registers and genres. Finally, I will introduce a new platform designed to help corpus linguists engage with the very latest AI models. This platform enables researchers not only to use AI to advance their own research and teaching but also to identify the strengths and weaknesses of current AI models and perhaps guide the development of the next generation of AI models.

PASCUAL PÉREZ-PAREDES

University of Murcia

Corpus linguistics and AI in the reconfiguration of language learning ecologies

AI is actively transforming language learning ecologies (Godwin-Jones, 2023). Rather than replacing traditional teaching, AI is increasingly part of a hybrid, co-constructed space where human and artificial intelligences interact to support learning. As Lévy (2025) has noted, AI represents a new phase in the evolution and manipulation of symbolic systems. They structure our reality. What we perceive, value, and understand is deeply shaped by the symbolic codes we inherit. In a postdigital world (Rowsell & Sandor, 2025), AI is not only creating new symbolic content, it is transforming meta-symbolic capabilities expanding the semiotic field and helping humans navigate symbolic complexity and reshaping how we interact with symbols.

AI is redefining teacher and learner roles, shifting from transmission to collaboration. It supports co-construction and reflective practices through dynamic interactions. Integrated into multimodal, socio-cognitive systems, AI enhances personalized learning and feedback. This evolution demands new literacies—critical, digital, and ethical—and reimagined pedagogical designs. As educators and learners adapt to the new technology (Godwin-Jones, 2021,2023), AI becomes a co-agent in learning, fostering environments where human and machine intelligence intersect to enrich the process of language acquisition. While AI presents challenges and ethical issues in education (McInnes, 2025), language educators and researchers have begun to consider ways into the co-creation and integration of large language models with the long-standing tradition of corpus-based approaches (Pérez-Paredes & Boulton, 2025).

This plenary explores the convergence of corpus linguistics (CL) and AI as complementary paradigms that, when combined, can offer a powerful framework for reimagining data-driven learning in second language education (Curry & McEnery, 2024; Pérez-Paredes & Boulton, 2025). Drawing on recent research from the Broadening the scope of Data-driven Language (BsDDL) project and practical models of AI and corpus literacy (Pérez-Paredes, 2024), I argue for an approach that foregrounds human agency, critical thinking, and metacognitive skill development. While AI affords accessibility and immediacy, CL anchors pedagogy in empirical, attested language use. Together, they support the design of pedagogies that are interactive, ethically grounded, and responsive to 21st-century learning goals. The talk discusses areas of convergence such as critical engagement, technological fluency, self-regulated learning, and interdisciplinary skill-building.

List of references

Curry, N., & McEnery, T. (2024). Corpus linguistics for language teaching and learning: A research agenda. Language Teaching, 1-20.

Godwin-Jones, R. (2021). Evolving technologies for language learning. Language, Learning & Technology, 25(3), 6–26.

Godwin-Jones, R. (2023). Emerging spaces for language learning: AI bots, ambient intelligence, and the metaverse. Language Learning & Technology, 27(2), 6-27.

Lévy, P. (2025). Symbolism, digital Culture and Artificial Intelligence. RED. Revista de Educación a Distancia, 25(81).

McInnes, R. (2025, April 11). Resist the gen-AI-driven university: A call for reclaiming thought in learning and teaching. ASCILITE TELall Blog. https://blog.ascilite.org/resist-the-gen-ai-driven-university-a-call-for-reclaiming-thought-in-learning-and-teaching/

Pérez-Paredes, P. (2024) Data-driven learning in informal contexts? Embracing Broad Data-driven learning (BDDL) research. In Crosthwaite, P. (Ed.). Corpora for Language Learning: Bridging the Research-Practice Divide, pp. 211-226. Routledge.

Pérez-Paredes, P. & Boulton, A. (2025). Data-driven Learning in and out of the Language Classroom. Cambridge University Press.

Rowsell, J., & Sandor, S. (2025). Literacy in Postdigital Times. In The Comfort of Screens: Literacy in Postdigital Times (pp. 26–45). Cambridge University Press.

PRE-CONFERENCE WORKSHOPS

Workshop 1.1: Reading concordances with algorithms

Nathan Dykes, Stephanie Evert, Michaela Mahlberg, Alexander Piperski

Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)

Context

Concordance analysis has long been central to corpus linguistics and other text-based disciplines, including digital humanities, computational social sciences, and computer-assisted language learning. It gives researchers a systematic lens for observing and interpreting patterns of language use, integrating both quantitative and qualitative perspectives. By focusing on a single search word or phrase in a context-limited display—commonly known as a KWIC (Key Word In Context)—scholars can investigate various aspects of its usage and meaning.

In spite of its wide applications, concordance reading has seen little innovation to date. Popular functions of concordance tools are still the traditional approaches, such as sorting lines alphabetically by the left or right context of the node or filtering for specific words. Another challenge for concordance reading is the documentation of the research process and methods applied, in order to ensure reproducibility. The workshop addresses these challenges by introducing both a taxonomy of concordance-reading strategies and a set of computational algorithms that build on these strategies to organize large amounts of textual data efficiently and transparently. Through hands-on exercises using the new Python library FlexiConc (https://pypi.org/project/FlexiConc/), which will be integrated into CLiC (https://clic-fiction.com/), the workshop will demonstrate how to apply robust concordance reading approaches to a variety of research contexts.

Aims

The workshop starts with an introduction to concordance analysis, including its place in the continuum of quantitative and qualitative research. We cover the most common general strategies for concordance analysis: selecting, sorting, and grouping lines, and show how each of them can aid interpretation. Participants will also learn about basic formal definitions and mathematical properties of the computational algorithms that underlie these strategies. We will discuss how algorithms extend beyond simple alphabetical ordering, opening up new possibilities for advanced text analysis.

The workshop will include a series of practical exercises, where participants first work with existing tools like AntConc or CQPweb and then proceed to exploring the functionalities of the FlexiConc library and its web interface. This library is designed to support a wide range of concordance reading strategies and to document user decisions in a systematic way. The web interface is designed to be intuitively accessible and to enable convenient interactive exploration. We introduce the concept of an ‘analysis tree’ to ensure the reproducibility and accountability of concordance research. By using a tree structure to trace the decisions taken when selecting lines from concordances, ordering, and grouping them, we can document not only the final results but also the process that led there. This approach fosters transparency, which is crucial for collaborative and interdisciplinary projects, as well as for replicating or extending research.

Format

Introduction to concordance analysis: fundamentals and strategies

Participants are introduced to basic concepts of concordance analysis. After a brief definition of fundamental terms and concepts, we give an overview of concordance software and its functionalities and allow participants to explore selected example concordances. Participants will be encouraged to share their observations on linguistic patterns as they work with existing concordancing tools.

We introduce strategies for organizing concordances (different types of selecting, ordering, and grouping). Each strategy is discussed with regards to its purpose, and how it may be combined with other strategies. In a hands-on exercise, participants apply different strategies themselves to example data and compare their observations to those from the step before to see how the application of dedicated strategies helps with concordance organisation and enhances systematicity.

Computational algorithms

Participants are introduced to our algorithmic approach to concordance reading, which extends the basic strategies and enhances their flexibility.

In a hands-on exercise, participants try out different concordance algorithms, including complex applications such as clustering, which are not widely available in current concordance tools. They can work with the web interface of our library on a public server, so no software installation is required (but advanced participants are welcome to work directly with the Python library).

Analysis trees for research documentation

We discuss reproducibility as a central challenge for concordance analysis and how this problem can be solved with the help of the 'analysis tree'. The tree-like display, accessible through the web interface of our library, enables users to trace and illustrate decision-making during concordance analysis.

Pre-requisites

This workshop targets an interdisciplinary audience, including students and researchers in corpus linguistics, general linguistics, computational linguistics, digital humanities, and computer-assisted language learning. We will keep the technical discussion to a manageable level to accommodate participants from both technical and non-technical backgrounds. Those interested in advanced techniques, such as more low-level concordance processing using Python, will be directed toward additional resources and follow-up materials after the session.

No software installation will be required prior to the workshop.

Workshop 1.2: Using ATLAS.ti for constructing and analysing multimodal social media corpora

Yuze Sha

Lancaster University

Context

Methods that enable comprehensive corpus analysis of multimodal data are essential for advancing our understanding of social media and digital communication. Social media posts are inherently multimodal, combining semiotic resources such as texts, emojis, memes, videos, and hyperlinks. Multimodal social media discourses have increasingly attracted research attention (Bouvier and Machin 2020; Djonov and Zhao 2013), whereas the methodological approaches remain largely divided into two camps: qualitative (e.g., Chałupnik and Brookes 2022; Hansson and Page 2023) and quantitative (e.g., Christiansen et al. 2020), each with strengths and limitations.

Within corpus linguistics, efforts to investigate multimodal discourse on social media are still at a relatively early stage. So far, no available corpus tool is capable of analysing multimodal data as systematically and comprehensively as monomodal linguistic data, such as by identifying (in)frequent co-occurrences and generating concordances across modes. In this workshop, I demonstrate how ATLAS.ti (version 25.0.1) can be used to construct and analyse multimodal social media corpora (Sha and Malory, in press). As a computer-assisted qualitative data analysis (CAQDAS) tool, ATLAS.ti is designed to support inductive, iterative data category development in line with grounded theory principles (Page 2022), offering transparent, systematic, and replicable workflows (Woods et al. 2016). Its quantifying functions, such as Code Co-occurrence Analysis and associated visualisation tools, can be flexibly applied to examine co-occurring patterns both within and across modes.

This workshop introduces four functionalities of the software that advance corpus-assisted discourse studies of multimodal social media data. It will be especially helpful for researchers who are interested in exploring the interplay between language and other modes of communication on social media platforms but currently lack dedicated tools for doing so.

Aims

This workshop has three primary aims.

- First, it will guide participants through the process of constructing a multimodal social media corpus using ATLAS.ti. This includes collecting and cleaning data from platforms such as Twitter, and addressing ethical considerations.
- Second, it will demonstrate how to use ATLAS.ti for problem-oriented corpus annotation and analysis, tailored to specific research questions. The session will cover annotation scheme design and introduce methods to: (1) obtain an overview of the corpus, (2) locate patterns of mono- and multimodal (non-)co-occurrences, (3) visualise these patterns, and (4) examine them in depth by reviewing the associated multimodal concordances and their (extra-)linguistic context.
- Third, the workshop will discuss the constraints involved in using ATLAS.ti for short-form multimodal social media research and propose practical strategies to address them.

By the end of the workshop, participants should have confidence in using ATLAS.ti to construct and examine their own multimodal social media corpora.

Format

Introduction to multimodal social media corpus design

- Overview of how multimodal corpora differ from linguistic monomodal corpora, as well as how multimodal social media corpora are distinct from other multimodal corpora.
- Ethical issues around collecting and using social media content.
- Considerations for data sampling, cleansing and annotation, with particular attention to the influence of social media affordances.

Data management in ATLAS.ti

- Demonstration of how to import and manage different semiotic resources, including texts, emojis, images, hyperlinks, and videos.
- Organising units of analysis (document groups, documents, quotations).
- Managing project files and backups.
- Guided hands-on practice.

Coding and techniques

- Developing annotation schemes aligned with specific research questions.
- Strategies for improving consistency and plausibility of annotation schemes.
- Guided hands-on practice.

Analysing and interpreting results

- Designing analytical procedures according to research questions.
- Generating quantitative insights using tools such as Code-Document Analysis, Code Co-Occurrence Analysis, and the Query Tool.
- Moving beyond surface patterns: integrating qualitative analysis of multimodal concordances and (extra-)linguistic features.
- Discussing how to combine various tools within ATLAS.ti to support comprehensive, research-driven analyses of multimodal social media corpora.
- Guided hands-on practice.

Pre-requisites

No prior experience with ATLAS.ti is required. However, familiarity with general concepts of corpus annotation and analysis would be helpful. Those who have worked primarily with monomodal linguistic corpora will benefit from learning how to include multimodal data, while those with qualitative multimodal research experience will gain insights into applying a more replicable, systematic approach to larger datasets.

ATLAS.ti is available in both desktop (Windows & macOS) and web versions, with the desktop version offering more advanced functionality. Participants are advised to bring their laptops with ATLAS.ti (version 22 or newer) pre-installed. Brief setup instructions will be provided ahead of the workshop and reiterated at the beginning of the session. Alternatively, the web version can be used for trial purposes, while it has some functional limitations.

List of references

Bouvier, Gwen, & David Machin. 2020. Critical discourse analysis and the challenges and opportunities of social media. Critical discourse studies and/in communication, 39-53.
Chałupnik, Małgorzata, & Gavin Brookes. 2022. Discursive acts of resistance: a multimodal critical discourse analysis of All-Poland Women's Strike's social media. Gender & Language, 16(3).
Christiansen, Alex, William Dance, & Alexander Wild. 2020. Constructing corpora from images and text. Corpus approaches to social media, 149-174.
Djonov, Emilia, & Sumin Zhao. 2013. From multimodal to critical multimodal studies through popular discourse. In Critical multimodal studies of popular discourse, 13-26.
Hansson, Sten, & Ruth Page. 2023. Legitimation in government social media communication: The case of the Brexit department. Critical Discourse Studies, 20(4), 361-378.
Page, Ruth. 2022. Analyzing multimodal interactions in social media contexts. Research Methods for Digital Discourse Analysis, 159.
Sha, Yuze, & Beth Malory. 2025. Using ATLAS.ti for constructing and analysing multimodal social media corpora. Linguistics Vanguard.
Woods, Megan, Trena Paulus, David Atkins, & Rob Macklin. 2016. Advancing qualitative research using qualitative data analysis software (QDAS)? Reviewing potential versus practice in

published studies using ATLAS. ti and NVivo, 1994–2013. Social science computer review, 34(5), 597-617.

Workshop 2.1: Open access and open source tools for corpus linguistics: Wmatrix version 7 and PyMUSAS

Paul Rayson, Daisy Lal, John Vidler, Andrew Moore

UCREL, Lancaster University

Context

This half day (3 hours) workshop will provide practical hands-on tutorial with the new version of the web-based Wmatrix corpus analysis and comparison software (https://ucrel.lancs.ac.uk/wmatrix/). Version 7 of Wmatrix is now open access for academic researchers and incorporates the Python open source (Apache Licence 2.0) version of the multilingual UCREL Semantic Analysis System (PyMUSAS) that automatically assigns semantic fields to words and multiword expressions to corpora (Rayson et al, 2004). Wmatrix7 via PyMUSAS provides support for 8 languages (https://pypi.org/project/pymusas/) and facilitates the extension of the key semantic domains method (Rayson, 2008) to those languages. Wmatrix7 represents the most significant update to the online software since the first version was presented at the ICAME 2001 conference (Louvain-la-Neuve, Belgium) and is now free to use. Wmatrix7 has a completely new indexing system implemented in the open source sqlite database allowing indexing of 10s of millions of words. The semantic lexicons used in PyMUSAS are also now freely available under Creative Commons CC-NC-BY-SA 4.0 licence (https://github.com/UCREL/Multilingual-USAS). Open access and open source tools are vital for the replicability and reproducibility of future corpus linguistics studies and support the explainability of annotation and analysis methods in corpus linguistics and NLP software, especially in light of the speedy uptake in new generative AI methods and large language models (LLMs), some of which are not open source or do not declare their training materials. Open tools also facilitate the exchange of methods and techniques to enable further developments to be built on top of existing groundwork e.g. as has been done in the Australian Text Analytics Platform (Jufri & Sun, 2022) building on PyMUSAS.

New and ongoing developments and features will also be highlighted including the future integration with large scale parallel processing using the UCREL-hex facility at Lancaster, a hybrid multiprocessor system including shared GPUs (https://www.lancaster.ac.uk/scc/research/research-facilities/hex/). Facilities like hex have been used to hugely speed up the large scale annotation of extreme scale corpora e.g. for the 1.2 billion words of the ParlaMint II corpus of comparable parliamentary data across Europe (Erjavec et al, 2024) from 18 days to around 7 hours. We will also describe further development of the English, Spanish, Dutch and Danish PyMUSAS taggers and lexicons as part of the 4D Picture project (https://4dpicture.eu/).

Aims

1. To provide a guided introduction to semantic annotation methods in corpus linguistics and natural language processing
2. To provide an introduction to the key semantic domains method and how it is operationalised in the Wmatrix7 tool along with PyMUSAS
3. To allow participants to explore the tools following guided tutorials and receive live direct feedback from the workshop organisers and tool developers themselves
4. Participants will also be given the opportunity to feed into future developments of the software via the collection of their requirements and preferences for new and adapted features in Wmatrix. They will also have the opportunity to discuss the development of PyMUSAS for new languages and to plan further collaborations.

Format

The workshop will begin with a 30 minute overview presentation introducing the theories and methods implemented in Wmatrix and PyMUSAS. The remainder of the time will be spent by participants being supported while following online tutorials to explore the tools using ready made corpora, e.g. the UK election manifestos corpora (https://github.com/perayson/manifestos) as well as to load in their own corpora for analysis.

Pre-requisites

Participants will likely have used other corpus linguistics software already, but the main methods (frequency lists, concordances, keywords, n-grams, collocations) will be introduced in the tutorials if needed. Participants with programming and command line experience will also be guided through the Python code necessary for use and integration of PyMUSAS in their own code, via Python Notebook demonstrators.

Wmatrix is a web based tool, so participants will only require an internet connection on their laptop or tablet plus a good web browser e.g. Chrome or Firefox. PyMUSAS will be demonstrated via web based access, but participants can bring their own laptops to install and run it locally using a Python programming environment, see https://pypi.org/project/pymusas/ for installation instructions.

List of references

Erjavec, T., Kopp, M., Ljubešić, N. et al. ParlaMint II: advancing comparable parliamentary corpora across Europe. Lang Resources & Evaluation (2024). https://doi.org/10.1007/s10579-024-09798-w

Jufri, Sony & Sun, Chao (2022). Semantic Tagger. v1.0. Australian Text Analytics Platform. Software. https://github.com/Australian-Text-Analytics-Platform/semantic-tagger

Rayson, P. (2008). From key words to key semantic domains. International Journal of Corpus Linguistics. 13 (4) 519-549. https://doi.org/10.1075/ijcl.13.4.06ray

Rayson, P., Archer, D., Piao, S. L., McEnery, T. (2004). The UCREL semantic analysis system. In proceedings of the workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks in association with 4th International Conference on Language Resources and Evaluation (LREC 2004), 25th May 2004, Lisbon, Portugal, pp. 7-12

Workshop 2.2: Word embeddings for discourse studies

Nathan Dykes[1], **Tim Feldmüller**[2]

[1]Friedrich-Alexander University of Erlangen-Nürnberg; [2]IDS Mannheim

Context

Word embeddings are vector representations of words. In order to generate them, each type in a corpus is transformed into a list of numbers through a neural network that has learned relationships between individual words and their syntagmatic contexts. This representation makes it possible to condense a vast spectrum of semantic (and morphosyntactic) information into the embeddings and to organize words in a vector space where words close to each other tend to share semantic and/or functional aspects (Bubenhofer 2020).

The analytical potential of word embeddings is complementary to that of collocations: embeddings reflect paradigmatic distributional similarity in the sense of "words that do not themselves co-occur, but whose surrounding words are often the same" (Sahlgren 2008: 43). The potential for Corpus Assisted Discourse Studies (CADS) is substantial: for instance, embeddings can help analysts to find (near-)synonyms that reflect prominent lexical fields. Training a model on a specialised target corpus can uncover associations that differ from typical language use in everyday discourse. This, in turn, can help identify covert attitudes and evaluations. For instance, one might investigate the distributional similarity of different person references to explore how similarly certain actors are represented.

While word embeddings and the language-model architectures built on them, such as BERT and GPT, have been ubiquitous in Computational Linguistics at least since the publication of Word2Vec (Mikolov et al. 2013), Corpus Linguistics and, in particular, CADS have shown little interest in word embeddings (see, however, for German e.g. Bubenhofer 2020; Knuchel & Bubenhofer 2023; Meier-Vieracker 2024). One important reason for this is that word embeddings are usually computed in programming languages like Python. However, extensive programming knowledge is not actually necessary for computing and carrying out basic analyses with word embeddings. Our workshop aims to fill this gap: We plan to provide the necessary skills to load and analyze both pre-trained word embedding models and to train word embeddings on one's own corpora.

Aims

Our workshop aims to provide an accessible introduction to applying word embeddings to Discourse Analysis. It is directed at researchers interested in applying corpus-driven quantitative methods to enable research questions where one quickly encounters limitations when using traditional Corpus Linguistic approaches. Applied CL, and CADS in particular, has often been focused on comparing frequencies on the level of individual words, making it challenging to explore themes realised through a wide range of lexical choices. Such examples may include phenomena tied to lexical fields (e.g. metaphor domains), near-synonyms or attitudes expressed with a wide variety of terms.

For the most part, analysts trained in linguistics explore corpora through tools that offer a graphical user interface. While these tools are convenient in terms of usability, they have limitations when it comes to incorporating more elaborate methods. At the same time, word embeddings are a valuable resource with significant potential for CADS.

While working in Python requires more introduction than a dedicated corpus platform, basic skills such as functions, variables and processing files can be quickly learned and transferable to various applications. Moreover, word embeddings as a specific resource are well-established in other fields and are available in relatively accessible formats. The skills conveyed in this workshop are thus intended to transfer relatively readily to participants' individual interests.

Format

Short introduction: "What are Word Embeddings?"

We begin with a concise introduction to the concept of word embeddings, explaining how words are represented as vectors and which potentials there are for discourse analysis.

Python basics (variables, functions, key data types, reading and writing files)

Participants receive an accessible overview of essential Python concepts for working with text. We work with prepared Jupyter Notebooks, allowing the participants to apply the concepts in an accessible environment.

Solving technical problems

We address typical technical issues that might occur during the hands-on sessions, such as installing libraries and setting up the computing environment.

Loading an existing model

We demonstrate how to load pre-trained word embedding models (one trained on a general language corpus and one on a more specialised thematic corpus) into Python, giving participants the opportunity to explore embedding spaces without the need to train a model from scratch.

Simple analyses (e.g. nearest neighbors, clustering)

Participants learn how to retrieve the nearest neighbors of a word in vector space and how to perform basic clustering or similarity analyses. They are encouraged to explore and reflect on the differences between the two provided models.

Training a custom model

We guide participants through the process of training a word embedding model on their own corpus (where available) or on sample data, highlighting important parameters and considerations.

Presentation of studies that use word embeddings

We showcase selected studies that apply word embeddings to discourse analysis. This includes hands-on discussion of examples from research, where participants are encouraged to critically engage with the results of embedding models and reflect on the potential for their own projects.

Discussion and further practice

If there is time left, participants can explore more advanced questions and discuss their own research interests.

Pre-requisites

No previous programming or in-depth technical knowledge is necessary. We will introduce all relevant concepts as we go along. Participants should have Jupyter Notebook installed on their computers before attending, as we will use it for all practical exercises. We will provide instructions for the installation before the workshop.

List of references

Bubenhofer, Noah. 2020. Semantische Äquivalenz in Geburtserzählungen: Anwendung von Word Embeddings. Zeitschrift für germanistische Linguistik 48(3). 562–589. https://doi.org/10.1515/zgl-2020-2014.

Knuchel, Daniel & Noah Bubenhofer. 2023. Machine Learning und Korpuspragmatik. Word Embeddings als Beispiel für einen kreativen Umgang mit NLP-Tools. In Simon Meier-Vieracker, Lars Bülow, Konstanze Marx & Robert Mroczynski (eds.), Digitale Pragmatik (Digitale Linguistik), 213–235. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-662-65373-9_10.

Meier-Vieracker, Simon. 2024. Racist Discourse in a German Far-Right Blog: A Corpus-Driven Approach Using Word Embeddings. Discourse & Society. SAGE Publications Ltd 35(2). 223–242. https://doi.org/10.1177/09579265231204510.

Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. arXiv. https://doi.org/10.48550/ARXIV.1301.3781.

Sahlgren, Magnus. 2008. The distributional hypothesis. Rivista die Linguistica 20(1). 33–53.

Workshop 2.3: Applying FAIR data principles in corpus linguistics

Iulianna van der Lek[1], Giulia Pedonese[2], Alexander König[1], Martin Wynne[3], Francesca Frontini[2], Megan Bushnell[3]

[1]CLARIN ERIC; [2]CNR-ILC; [3]University of Oxford

Context

Recent projects and initiatives acknowledged that there is a lack of general awareness among lecturers, students and researchers of research data management practices, including knowledge of the FAIR data principles (https://www.go-fair.org/fair-principles/) to make digital resources more Findable, Accessible, Interoperable and Reusable (e.g. UPSKILLS project (https://upskillsproject.eu/), EOSC Skills and Training Working Group (https://eosc.eu/opportunity-area-exp/oa5-skills-training-rewards-recognition-upscaling/), EC Digital Skills for FAIR and Open Science (https://data.europa.eu/doi/10.2777/59065)).

Therefore, the FAIR Competence Framework for Higher Education proposes a set of core competencies for FAIR data education that universities can use to design and integrate research data management and FAIR-data-related skills in their curricula and programmes (Demchenko et al., 2021). Students, scholars, teachers and researchers from all disciplines are encouraged to acquire fundamental skills for open science, including the ability to effectively interact with federated research infrastructures and open science tools for collaborative research. To further support the integration of these skills into the university curricula, an adoption handbook was published "How to be FAIR with your data – A teaching and training handbook for higher education institutions" (Engelhardt et al., 2022, https://fairsfair.gitbook.io/fair-teaching-handbook), which contains ready-made lesson plans on a variety of topics, including the use of repositories, data creation and reuse. In addition, the Skills4EOSC project (https://www.skills4eosc.eu/) provides an adaptable framework focusing on digital skills and using existing technologies to improve the competencies and skills of researchers. Specifically for linguistics, and other humanities disciplines, teaching resources and best practices guidelines have been created in the UPSKILLS and H2IOSC projects (Degl'Innocenti et al., 2023) to show various target audiences how the CLARIN research infrastructure (https://www.clarin.eu/) can support researchers in adopting and applying the FAIR data principles in their research practices. Based on the experience acquired in these projects, the authors of the abstract propose a workshop to raise awareness of the FAIR (and, to a lesser degree, the CARE principles for Indigenous Data Governance (https://www.gida-global.org/care)) and how they can be taken as guidance in corpus linguistics projects to ensure that the language research data is not only FAIR but also follows ethical research practices and supports Open Science. Hands-on demonstrations will be included using services, tools and language resources from the CLARIN research infrastructure.

Aims

This workshop will show participants how to incorporate the FAIR and CARE (Compared to the more generally applicable FAIR principles, the CARE principles focus specifically on certain research scenarios and will therefore play a less prominent role in the workshop.) principles into their corpus linguistics research projects. The programme will consist of theoretical and hands-on exercises, including services and tools from CLARIN, a European Research Infrastructure for language as social and cultural data. Through a combination of theoretical principles, hands-on activities and case studies, the participants will learn how to identify the requirements for a linguistic resource (e.g. a linguistic corpus) to align with the FAIR and CARE principles and apply them in their research workflow. Finally, the workshop will contain a case study and a roleplay on how to write a Data Management Plan (DMP) for your research. The case study will focus on an early-career researcher's experience working on a research project in corpus linguistics. The workshop participants will learn how to draft a DMP using a sample research project as an example and get to know the Argos application (https://argos.openaire.eu/home). This tool, developed by OpenAIRE, allows scholars to write, save and export their DMP according to FAIR principles and Open Access best practices.

By the end of this workshop, the workshop participants will be able to:

- Identify the requirements for a resource to align with the FAIR and CARE principles
- Find and use certified research data repositories for data collection, sharing and archiving

- Create the outline of a research data management plan and familiarise themselves with the Argos application
- Identify and use infrastructure tools for data processing and analysis

Format

Introduction

- What is CLARIN?
- What are FAIR and CARE principles, and how can they be applied in corpus linguistics?

Finding and analysing linguistic resources in CLARIN

- How CLARIN supports the FAIRness of data
- Guided tour of CLARIN's language data discovery portal, the Virtual Language Observatory (https://vlo.clarin.eu)
- Tool examples from the Language Resource Switchboard (https://switchboard.clarin.eu/): processing a text with Weblicht (demo)

Creating a Data Management Plan (case study and discussion)

Demo of depositing, sharing and archiving your corpus data

Pre-requisites

No previous knowledge of FAIR and CARE principles is required.

Institutional login to access the CLARIN services (Most academic accounts can be used for logging into CLARIN services thanks to the CLARIN service provider federation, see https://www.clarin.eu/content/federated-identity for details). Please test beforehand. If you encounter access issues, you can request a CLARIN account at https://user.clarin.eu/user/register.

List of references

Schulder, Marc and Hanke, Thomas. (2022). "How to be FAIR when you CARE: The DGS Corpus as a Case Study of Open Science Resources for Minority Languages". In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 164–173, Marseille, France. European Language Resources Association.

Mark D. Wilkinson et al., (2016). The FAIR guiding principles for scientific data management and stewardship, Scientific Data3, Nr. 1: 160018, https://doi.org/10.1038/sdata.2016.18

Mattern, Eleanor (2022). "The Linguistic Data Life Cycle, Sustainability of Data, and Principles of Solid Data Management", The Open Handbook of Linguistic Data Management, Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister.

van der Lek, I., Fišer, D., Samardzic, T., Simonovic, M., Assimakopoulos, S., Bernardini, S., Milicevic Petrovic, M., & Puskas, G. (2023). Integrating research infrastructures into teaching: Recommendations and best practices (Version 2). Zenodo. https://doi.org/10.5281/zenodo.8114407

van der Lek, Iulianna; Fišer, Darja. (2023). Introduction to Language Data: Standards and Repositories. In UPSKILLS Learning Content. https://upskillsproject.eu/project/standards_repositories/. CC BY 4.0.

Degl'Innocenti, Emiliano, Monica Monachini, Alberto Bucciero, Enrico Pasini, Bruno Fanini, e Francesca Frontini. «H2IOSC: Humanities and Heritage Open Science Cloud». In La memoria digitale: forme del testo e organizzazione della conoscenza. Atti del XII Convegno Annuale AIUCD, a cura di Emmanuela Carbé, Gabriele Lo Piccolo, Alessia Valenti, e Francesco Stella, 63–64, 2023. https://iris.unive.it/retrieve/0f226d38-e332-418b-9b14-d5558d1a0d9d/AIUCD2023.pdf.

THEMATIC PANELS

MONDAY 30th JUNE

Methodological innovation in applied corpus linguistics (BAAL Corpus SIG panel)

Robbie Love[1], Siân Alsop[2], Debora Cabral[3], David Griffin[3], Dana Roemling[4]

[1]Aston University; [2]Coventry University; [3]Cardiff University; [4]University of Birmingham

This panel brings together three studies discussing innovative approaches to addressing both longstanding and emergent methodological challenges in applied corpus research. Specifically, the presentations discuss challenges related to corpus compilation, keyness analysis, and the analysis of multimodal data.

The panel is associated with the BAAL (British Association for Applied Linguistics) Corpus Linguistics Special Interest Group (SIG).

Story words: Creating a large corpus of children's story books

Siân Alsop

This paper discusses methodological challenges in the creation of the Story Book Corpus, a large corpus of 1000 early years story books. The corpus will be used to inform a classroom vocabulary acquisition intervention by identifying words that are highly characteristic of story books, in frequent usage, and sensitive to vocabulary change in children aged 3-5 years old. The Story Book Corpus is representative of texts that are commonly available in early years settings across the UK. Sampling was informed by national supplier data, library usage data, national and local nursery recommended reading lists and audits of current on-shelf holdings (local nurseries and individual homes). Corpus compilation involves scanning physical books and converting story text to a machine-readable form, then associating this 'body' content with extensive metadata about the text and author to create corpus files. We used keyword analysis techniques to identify the words most characteristic of these story texts by comparing our corpus to an age-appropriate subcomponent of the CHILDES corpus, via SketchEngine. This paper discusses practical elements of creating a large corpus of this nature, and various aspects of decision-making required to whittle down a list of over 1000 items to 40 testable words. Its story is one of a messy journey with a happy ending.

Relative Frequency and RF:ARF ratio: Addressing the challenge of combining relevance and dispersion in keyness analysis

Debora Cabral

This paper proposes a new approach to keyness analysis which accounts for both the keywords' relevance and dispersion. Keywords can be measured using effect-size and/or statistical significance metrics (Gabrielatos, 2018: 232); the limitations of using one or the other have been discussed variously (Baker, 2004; Egbert & Biber, 2019; Gabrielatos, 2018: 243; Gries, 2008; Kilgarriff, 2009). This paper proposes a combination of the keywords' relative frequency with the ratio between their raw frequency and average reduced frequency (RF:ARF ratio; Savicky & Hlavácová, 2002) in order to identify the linguistic realisation of activism. The keyword analysis compares an Activism corpus (of 474 news posts from WWF-Brasil and 474 news posts from Greenpeace (Brazil), amounting to 948 texts and 682,846 tokens) with Sketch Engine's Portuguese Web 2018 (ptTenTen18) (Kilgariff et al., 2014). After identifying the most frequent and most evenly distributed keywords in the Activism corpus and excluding the environmental-activism related keywords, the remaining keywords comprise a preliminary activism lexicon. Analysis of co-text provides evidence of the linguistic realisation of activism through patterns of "opposition", "engagement" and the justification of change for the "good of society". These findings support the theoretical assumption that discourse structures materialise the individuals' cognitive perception of the world (van Dijk, 2018), and future research can identify how other social representations are linguistically realised.

Looking beyond lexis: A corpus semiotic approach to the analysis of static texts

David Griffin, Dana Roemling

The application of corpus linguistic methodologies to multimodal texts is a complex but worthwhile endeavor. By considering the full range of semiotic resources deployed across a corpus (e.g. the use of color and choice of typeface in a series of technical manuals), researchers are able to gain insight into data which would not be accessible by considering word choice alone. This paper discusses strategies for the effective semiotic analysis of static multimodal texts, including the use of frameworks established specifically for this purpose (e.g. Bateman, 2008) and methods for adapting techniques designed with other contexts in mind (e.g. Le Roux & Rouanet, 2010). Particular attention is paid to the effective presentation of the results of such analyses via the use of heatmaps (Griffin & Roemling, 2024) and other visualization approaches. Sample analyses of a corpus of legal texts will be shown and discussed.

Following the paper presentations, the final slot will comprise an interactive roundtable session discussing specific questions arising from the papers as well as a broader discussion of emergent methodological challenges in applied corpus research.

List of references

Baker, P. (2004). Querying keywords: Questions of difference, frequency, and sense in keyword analysis. Journal of English Linguistics 32(4), 346-359.

Bateman, J. A. (2008). Multimodality and genre: a foundation for the systematic analysis of multimodal documents. Palgrave Macmillan.

Crosthwaire, P., & Baisa, V. (2023). Generative AI and the end of corpus-assisted data-driven learning? Not so fast! Applied Corpus Linguistics, 3(3), 1-4.

Curry, N., Baker, P., & Brookes, G. (2024). Generative AI for corpus approaches to discourse studies: A critical evaluation of ChatGPT. Applied Corpus Linguistics, 4(1).

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

Gabrielatos, C. (2018). Keyness analysis: Nature, metrics and techniques. In C. Taylor and A. Marchi (eds). Corpus Approaches to Discourse: A critical review. Routledge, 228-258.

Gries, Th. G. (2008). Dispersion and adjusted frequencies in corpora. International Journal of Corpus Linguistics, 13(4), 403-437.

Griffin, D., & Roemling, D. (2024). Signs of Legal and Pseudolegal Authority: A Corpus-based Comparison of Contemporary Courtroom Filings. International Journal for the Semiotics of Law. https://doi.org/10.1007/s11196-024-10183-7

Kilgarriff, A. (2009). Simple maths for keywords. In M. Mahlberg, V. González-Diaz, C. Smith (eds.) Proceedings of Corpus Linguistics Conference 2009, University of Liverpool.

Kilgariff, A., Baisa, V., Busta, J., Jakubicek, M., Kovar, V., Michelfeit, J., Rychly, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. Lexicography, 1, 7-36.

Le Roux, B., & Rouanet, H. (2010). Multiple Correspondence Analysis. SAGE.

Savicky, P., & Hlavácová, J. (2002). Measures of word commonness. Journal of Quantitative Linguistics, 9(3), 215-231.

Uchida, S. (2024). Using early LLMs for corpus linguistics: Examining ChatGPT's potential and limitations. Applied Corpus Linguistics, 100089, https://doi.org/10.1016/j.acorp.2024.100089

Linguistic and identity dynamics: A corpus-driven analysis of the specialised discourse of *drag*

Macarena Palma Gutiérrez[1], Eva Lucia Jimenez-Navarro[1], Beatriz Martín-Gascón[2], Mariola Ruiz Rodríguez[1], M. Azahara Veroz-Gonzalez[1], Nicolás Müller Pastene[2]

[1]Universidad de Córdoba; [2]Universidad Complutense de Madrid

Despite the increasing visibility of drag culture in mainstream media (Toquero 2018; Fusco 2020; Taylor 2020a; Crookston 2021), the linguistic features of the drag discourse remain underexplored in corpus-based studies. This research addresses this gap by analysing the specialised language of drag using a dedicated corpus derived from the original subtitles of the talent show RuPaul's Drag Race. By using the Sketch Engine software, this panel explores aesthetics and linguistics on 'camp' (Simple 2022), that is, it examines how drag performers employ language to construct identity, challenge societal norms, and influence conventional discourse (Fonseca and Quintero 2009; Taylor 2020b; Taylor 2021).

Based on our corpus analysis, this roundtable will explore three key areas of the language of drag: (1) its lexico-grammatical features, (2) relevant cognitive and cultural models in drag discourse, and (3) the role of queer identity and linguistic creativity in subverting established norms.

Panel Structure and Duration

This panel is structured into three 30-minute blocks (totalling 1.5 hours). Each block consists of a 25-minute presentation followed by 5 minutes for questions and answers.

Block 1: Lexico-Grammatical Characterisation of the Specialised Language of Drag

Macarena Palma Gutiérrez, Eva Lucia Jimenez-Navarro

A distinctive feature of the language of drag is its rich lexico-grammatical structure, which includes neologisms, diverse morpho-syntactic processes, collocations, and idiomatic expressions, that is, linguistic processes that contribute to the playful, boundary-pushing nature of the discourse. For example, drag culture is known for its creative neologisms (such as eleganza extravaganza, describing something extremely elegant and extravagant) and its morphological flexibility (as illustrated by condragulations, a playful combination of "congratulations" and "drag"). Collocations and idiomatic expressions also characterise this language, with expressions like read for filth (in other words, to deliver a sharp critique) becoming widely known. These linguistic innovations not only reflect the performative nature of drag but also underscore its ability to reshape language in line with its values of transformation and self-expression.

Block 2: Cognitive and Cultural Models in the Drag Community

Beatriz Martín-Gascón, Nicolás Müller Pastene

The language of drag is also shaped by cognitive and cultural models that reflect values, structures, and experiences within the drag community. A key aspect of this is the use of metaphors and metonymies, which are central to how drag queens construct their identities, and thus, language. Cognitive models of motherhood and royalty are particularly influential in drag culture, offering frameworks for understanding relationships, power, and performance. Motherhood refers to the nurturing and mentoring relationships between queens, often embodied through the concept of drag families, where older queens (the "mothers") mentor younger queens (the "children"). This metaphor reflects the emotional and supportive bonds that are central to the drag community. Similarly, royalty serves as a powerful motif, with queens positioning themselves as "royal" figures in a performance of authority and grandeur. This is reflected in expressions like house (involving the metonymic formula place for people to refer to the belonging to a powerful drag dynasty). These expressions not only highlight the performative nature of drag but also speak to the status and power that drag queens cultivate through their art.

Intertextual references further enrich the drag language. Queens often draw on cultural icons and queer history, making frequent references to figures like Judy Garland, whose legacy is honoured in the term judy (meaning "best friend"). These references allow drag performers to weave a sense of continuity and cultural pride into their performances, linking the past and present of queer culture. Additionally, formulaic expressions such as RuPaul's signature catchphrase, Good luck, and don't fuck it up,

emphasise competition, transformation, and self-assurance—values that are central to both drag performance and the broader cultural impact of the show.

Block 3: Queer Identity and Linguistic Creativity as Subversion

Mariola Ruiz Rodríguez, M. Azahara Veroz-Gonzalez

The language used by drag queens plays a critical role in the expression of queer identities, offering a form of linguistic creativity that challenges the repression of gender and sexual diversity. This language often embodies the humour, wit, and sharpness central to drag culture and frequently enters real-world/overall/ordinary usage. Historically, forms of secretive slang, such as Polari (Baker 2002), have allowed queer communities to express themselves in the face of societal persecution. In contemporary drag culture, linguistic forms like euphemisms and slang (e.g., beat for flawless makeup or shade for subtle insults) continue to serve as tools of resistance and empowerment. These expressions allow drag queens to critique and subvert traditional norms of gender, beauty, and performance, often with humour and irony.

The fan-based internet subculture surrounding drag, particularly through platforms like Twitter, Instagram, and TikTok, has been crucial in spreading the language of drag to a broader audience. The sharing of memes, GIFs, and videos enables the linguistic innovations of the drag community to permeate popular culture, illustrating how fandoms create cultural bridges between subcultures and the general public. This widespread dissemination of drag-related media has contributed to the normalisation of the queer language and its growing presence in everyday discourse.

Conclusion

This roundtable seeks to offer a comprehensive exploration of the linguistic, cultural, and identity-related aspects of the drag language by relying on the data extracted from a specialised corpus built ad hoc. By examining its lexico-grammatical features, the cognitive and cultural models that inform its use, and its role in expressing and subverting queer identity, we aim to highlight how this type of discourse is not only a tool for individual expression but also a force that shapes mainstream cultural norms. As the language of drag moves from subculture to the wider public, it brings with it new understandings of gender, performance, and identity. Through the lens of the drag discourse, we can see how language functions as both a mode of resistance and a means of empowerment, facilitating a deeper understanding of the cultural dynamics of queerness and belonging in the 21st century. By analysing these linguistic phenomena, we hope to enrich both academic discourse and public recognition of the drag culture's role in redefining norms of gender and identity.

List of references

Baker, P. (2002). Polari - The lost language of gay men. Routledge.

Crookston, C. (2021). Introduction: Why are we all gagging? Unpacking the cultural impact of RuPaul's Drag Race. In C. Crookston (Ed.), The cultural impact of RuPaul's Drag Race. Why are we all gagging? (pp. 1-10). Intellect.

Fonseca, C., & Quintero, M. L. (2009). La teoría queer: la de-construcción de las sexualidades periféricas. Sociológica, 24(69), 43-60.

Fusco, M. P. (2020). RuPaul's Drag Race en francais. The influences of modern LGBTQ media translation on queer identity and visibility. (Master's Dissertation). Concordia University Press.

Simple, K. (2022). Camp Studies, and Queer Theory, and drag queens, oh my! Camping the academy, queer methods, and the potentiality of camp. (Thesis Dissertation). The Australian National University.

Taylor, A. S. (2021). Repetition, recitation, and Vanessa Vanjie Mateo: Miss Vanjie and the culture-producing power of performative speech in RuPaul's Drag Race. In C. Crookston (Ed.), The cultural impact of RuPaul's Drag Race. Why are we all gagging? (pp. 175-193). Intellect.

Taylor, A. S. (2020a). Repetition, remix and reproduction: memes as visual deconstruction. In A. S. Taylor (Ed.), Authenticity as performativity on social media (pp. 109-132). Springer.

Taylor, A. S. (2020b). Authentic self-representation. In A. S. Taylor (Ed.), Authenticity as performativity on social media (pp. 51-79). Springer.

Toquero, M. M. (2018). El surgimento de las drag Queens, una forma de expresión que se populariza entre la comunidad LGBT. Epikea, Revista del Departamento de Ciencias Sociales y Humanidades, 35, 1-12.

TUESDAY 1st JULY

Analysing conversational American English: Studies on the Lancaster-Northern Arizona Corpus of American Spoken English (LANA-CASE)

Paul Baker[1], Douglas Biber[2], Raffaella Bottini[1], Vaclav Brezina[1], Gavin Brookes[1], Isobelle Clarke[1], Rachele De Felice[3], Jesse Egbert[2], Elizabeth Hanks[2], Alexander Holmberg[2], Taehyeong Kim[2], Tove Larsson[2], Jacqueline Laws[4], Tony McEnery[1], Nele Põldvere[5], Randi Reppen[2]

[1]Lancaster University; [2]Northern Arizona University; [3]Open University; [4]University of Reading; [5]Lund University

This thematic panel introduces a new corpus of American English conversation: the Lancaster-Northern Arizona Corpus of American Spoken English (LANA-CASE). This corpus is an open-access, large-scale collection of spoken American English which is scheduled to be released in 2026. The proposed panel will present LANA-CASE (design, sampling and representativeness) and some empirical studies that utilize a 5-million-word subset of the data to (a) provide insights into conversation language in the U.S., (b) highlight potential applications of this corpus through a variety of methodological approaches, and (c) inform future research about both strengths and limitations of the corpus.

The following sections describe each block of the panel, where the first block provides information about the corpus, and following blocks present studies that utilize a subset of LANA-CASE data and discuss their implications.

Introducing the Lancaster-Northern Arizona Corpus of American Spoken English (LANA-CASE)

Elizabeth Hanks, Tony McEnery, Jesse Egbert, Tove Larsson, Alexander Holmberg, Douglas Biber, Randi Reppen, Paul Baker, Vaclav Brezina, Gavin Brookes, Isobelle Clarke, Raffaella Bottini

The Lancaster-Northern Arizona Corpus of American Spoken English (LANA-CASE) consists of 10 million words of everyday, unscripted, spoken American English from speakers of different regions, ages, genders, and race/ethnicities. In this presentation, we describe corpus design and sampling, especially those elements which most strongly impact the content of the corpus (i.e., utilization of public participation in scientific research, following Love et al., 2017). We then discuss the proportional representation of demographic characteristics (e.g., region, age, gender, race/ethnicity) and situational variables (e.g., relationship between interlocutors, communicative purposes) within LANA-CASE. Finally, we provide suggestions for use, including availability of the data, ethical considerations of corpus use, and limitations to consider in future analyses.

Is conversation linguistically distinct from other spoken registers?

Tove Larsson, Douglas Biber, Taehyeong Kim

We know from previous research that there is limited variability among spoken registers when it comes to the frequency of use of grammatical complexity features (e.g., attributive adjectives; see Larsson et al., 2024). This has been attributed to limitations of spoken processing and production (Biber, 1992). Given these findings, we are left wondering if it really matters what spoken register we study. If we are interested in spontaneous conversation, would it work equally well to study interviews?

The answer might be 'yes' for grammatical complexity; however, other linguistic features may vary in terms of how sensitive they are to different situational characteristics. The present paper looks at four spoken registers that differ across situational characteristics such as interactivity and preparedness. In addition to grammatical complexity features, we look at three sets of linguistic features that we expect would be affected by different kinds of situational characteristics: interactional features, stance features, and structurally reduced features.

Advice in American English conversation: A corpus pragmatics study

Nele Põldvere, Rachele De Felice

Advice is an important, yet sensitive, social action in spontaneous conversation. In a previous corpus study based on British English conversation (Põldvere, De Felice & Paradis, 2022), we established the range of constructions that advisers use to express communicative (in)directness to different degrees, and that *who* and *how* advice is given are the strongest predictors of advice uptake. In the present study,

we seek to investigate whether these patterns are the same in American English conversation as well as to extend the analysis to investigate a subset of advice constructions that have interesting persuasive qualities in discourse, namely, indefinite pronouns (*everyone puts milk in their coffee*). Our analysis will focus particularly on the socio-demographic variables of 'age' and 'gender' to match the categories in Põldvere, De Felice and Paradis (2022).

…and I was like wait what? The Language of Gen Z Americans

Paul Baker

In the absence of a diachronic corpus, an analysis of age variation can provide clues with regard to potential changes in language use over time. Earlier studies of age variation in the two iterations of the spoken BNC have noted age grading patterns in use of *may* as a modal verb (Baker & Heritage, 2021), or greater use of hyperbole among young people (Collins & Baker, 2023). This study uses #LancsBox to examine distinctive language use by members of Gen Z (born 1997-2012) in LANA-CASE, a group characterized by high use of technology, elevated levels of depression, anxiety and individualism and tolerance of diversity. The analysis focusses on key words, part of speech tags and n-grams. A set of features associated with Gen Z were then compared against features derived from other spoken corpora to identify whether such features are simply typical of 'young person' style or if tell us something unique about Gen Z.

Neologizing in American and British spoken English: The case of complex verb formation

Jacqueline Laws

Complex verb formation in English involves the packaging of concepts into a single word, predominantly through the attachment of four verb-forming suffixes: *-ize, -ify, -en* and *-ate* (e.g., *trivialize*, 'to regard something as trivial', and *hyphenate*, 'to insert a hyphen between words'). Currently, little is documented on novel complex verbs (e.g., *yogarize*, 'to practise yoga') in American and British English. Laws (in press) found that the density of complex verb neologisms in COCA spoken texts for the periods 1990-1994 and 2010-2014 was lower than that observed in their time-matched British spoken counterparts, BNC1994 and BNC2014. In addition, Laws et al. (2017) noted that in the conversational BNC1994 sub-corpus (DS1994) more complex verb neologisms were coined by females than males. This paper examines this under-explored area by analysing the forms, meanings and relative frequency of newly-coined complex verbs occurring in everyday American (LANA-CASE) and British speech (Spoken BNC2014), as a function of speaker education, age and gender.

Panel discussion and questions

Led by Tony McEnery

Panelists respond to audience questions about the studies and the corpus itself. Time allowing, panelists may also discuss the following questions:

1. How do you envision LANA-CASE contributing to future linguistic research?
2. How might corpus users balance generalizability through analyses of the corpus as a whole with the richness of individual conversation files?
3. Do you anticipate any technological developments that could further enhance the utility of LANA-CASE for future research?
4. What advice do you have for researchers interested in using the corpus for sociolinguistic studies?

List of references

Biber, D. (2003). Variation among university spoken and written registers: A new multi-dimensional analysis. In Corpus analysis (pp. 47-70). Brill.

Larsson, T., Biber, D., & Hancock, G. R. (2024). On the role of cumulative knowledge building and specific hypotheses: The case of grammatical complexity. Corpora, 19(3), 263-284.

Laws, J. (in press). The role of coercion in the productivity and creativity of complex verb formation: A constructional approach. English Language & Linguistics.

Laws, J. Ryder, C. and S. Jaworska (2017). A diachronic corpus-based study into the effects of age and gender on the usage patterns of verb-forming suffixation in spoken British English. International Journal of Corpus Linguistics, 22(3), 375-402.

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.

Põldvere, N., De Felice, R., & Paradis, C. (2022). Advice in conversation: Corpus pragmatics meets mixed methods. Cambridge University Press.

WEDNESDAY 2nd JULY

Celtic languages in the digital age: Latest developments

Beatrice Alex[1], **Megan Bushnell**[2], **Mo El-Haj**[3], **Dawn Knight**[4], **Will Lamb**[1], **Mícheál J. Ó Meachair**[5], **Paul Rayson**[3], **Martin Wynne**[2]

[1]University of Edinburgh; [2]University of Oxford; [3]Lancaster University; [4]Cardiff University; [5]Dublin City University

The panel session will be structured with a short introduction from the chair(s), and then three papers, each offering an overview and focussing on the latest development in digital resources in one of the main Celtic languages: Irish, Welsh and Scottish Gaelic. Discussion will also invite questions and the points relating to other Celtic languages and related matters.

Chairs: Megan Bushnell, Martin Wynne

Paper 1: From Sparse Data to Large Language Models: The Evolution and Future of Scottish Gaelic Language Technology

Beatrice Alex, **Will Lamb**

Dr. Beatrice Alex, Senior Lecturer and Chancellor's Fellow in Text Mining and Professor Will Lamb in Gaelic Ethnology and Linguistics at University of Edinburgh have played a pivotal role in advancing Scottish Gaelic language technology. Initially, the lack of digitised Gaelic data posed significant challenges, classifying it as a low-resource language in natural language processing(NLP). In this talk, Alex and Lamb will present collaborative efforts to develop Gaelic handwriting recognition, speech recognition and text transformation tools aimed at increasing the availability of digitised Gaelic texts for NLP model training. They will also share insights into an ongoing project on digitising and analysing Gaelic and Irish folktales. Looking ahead, they will outline future work planned focusing on how we can responsibly develop advanced language technologies using large language models for speakers of this minority language.

Paper 2: Digital Resources for the Irish Language - latest developments

Mícheál J. Ó Meachair

In this panel discussion Dr Mícheál J. Ó Meachair presents the newly launched National Corpus of Irish (CNG) website, corpas.ie. The presented work was completed by the Gaois research group, Fiontar & Scoil na Gaeilge, DCU as part of a three-year research project with funding from the Department for the Gaeltacht and the National Lottery. The project website utilises a mixture of open-source and newly-created tools and technologies in order to provide a corpus tool that is accessible to, and easily-used by, lay people and experts alike. These tools and technologies include a newly-developed part-of-speech tagger for modern Irish, an instance of NoSketchEngine, a word-embeddings search feature, an interpreter that functions as a POS-tag generator, and a word-frequency navigator. This demonstration of tools and features is accompanied by use-cases from our user base, and is followed by a description of the planned additions and developments.

Dr Ó Meachair's presentation concludes with examples of how this corpus resource is being used to assist the second-language acquisition (SLA) process across levels of ability. The project team believes it is very important that smaller languages or minority languages have this type of support. Census data indicates Irish is maintaining its native and fluent daily-speaker population, but the largest cohort of Irish speakers in Ireland is still our learner community. Supporting these learners on their journey to becoming fluent daily speakers is, therefore, as important now as ever.

Paper 3: Digital Resources for the Welsh Language - latest developments

Dawn Knight, **Paul Rayson**

The creation of language technology resources in a minoritised language context, such as the Welsh language, poses interesting challenges, but also presents opportunities that are not always available to developers of such resources for larger languages. In this presentation we demonstrate how scrutiny of the unique context of a specific minoritised language, and meaningful collaboration with potential user groups, can determine the design and construction of language resources. This paper showcases recent developments in the creation of digital resources for Welsh.

This showcase will include a discussion of some of our own interdisciplinary and cross-institutional projects including CorCenCC corpus (the National Corpus of Contemporary Welsh: www.corcencc.org), Thesawrws, FreeTxt (a bilingual toolkit that supports the analysis and visualisation of free text data), GDC-WDG (an online collection of freely available digital resources designed to support the exploration, analysis, learning, and referencing of the Welsh language) and the development of a small language model for Welsh. The creation of these resources involved the development of important new tools and processes, including, in the case of CorCenCC, a unique user-driven corpus design in which language data was collected and validated through crowdsourcing, and an in-built pedagogic toolkit (Y Tiwtiadur) developed in consultation with representatives of all anticipated academic and community user groups. The approaches used to construct the resources mentioned in this talk provide an invaluable template for those researching other minoritised languages. The specifics of how this template might inform resource development in Welsh and other languages will be discussed further during the presentation.

Regulated discourse: Formal records of debates and their generic constraints

Sascha Diwersy[1], Hugo Dumoulin[2], David Kahn[3], Giancarlo Luxardo[1], Cyirlle Montrichard[4], Timothée Premat[4], Frédérique Sitri[4]

[1]Praxiling UMR 5267, CNRS - Université de Montpellier Paul-Valéry; [2]MoDyCo UMR 7113, CNRS - Université Paris Nanterre; [3]FRAMESPA UMR 5136, CNRS - Institut National Universitaire Champollion; [4]CEDITEC, Université Paris-Est Créteil

Outline

When analysing the formal records of discussions produced in diverse institutional contexts, one might expect to see ideology explicitly driving the issues at stake, with debates transparently reflecting these influences. However, this is rarely the case: such records include a range of formulaic expressions, shaped by the conventions and frameworks governing institutions such as parliamentary assemblies, university councils, or judicial bodies. These frameworks operate within what might be described as "institutional encapsulation" (*clôture institutionnelle*), where discourse is tightly constrained by the rituals and operational norms of the institution. This phenomenon is closely tied to the process of delegation that defines the political sphere, creating a divide between institutional agents—through whom, as Bourdieu (1982: 213) notes, "politically active and legitimate forms of perception and expression are constituted"—and the wider, everyday sphere of their constituents.

We hypothesise that this phenomenon of institutional encapsulation is accompanied by a discursive shaping process which manifests through specific genres of discourse that participate in the 'preconstruction' of speech (Bakhtin, 1984). Based on corpora from various institutional contexts, we explore how corpus linguistics methods can illuminate the ways in which formal records of discussions are shaped by institutional constraints.

The contributions examine the challenges involved in deploying such methods, particularly in describing the processes through which formulaic and preconstructed expressions take shape within discursive routines. The panel will feature five talks and a round table.

Assessing political divides in French parliamentary debates

Designed to meet the requirements of researchers from several disciplines, ParlaMint (Erjavec et al. 2024) is a resource providing parliamentary records in large amounts, in multiple languages and originating in multiple countries. The corpus includes extensive metadata describing actors, mandates, and organizations (political bodies) referenced in the proceedings.

This contribution is based on the French corpus, which will be used in the study of a controversial government bill debated in 2021. When the objective is to recognize and discriminate political ideologies in the debates, the results of a textometric analysis reveal a significant “discursive blending”, which makes this objective challenging. This phenomenon is due to the presence of an institutional vocabulary, characteristic of a parliamentary protocol, strictly codifying the interactions among different types of actors. As a result, every actor should be considered, besides political affiliation with a constrained role as a speaker in the debate.

Discursive Routines and Speakers Status in University Councils' Minutes

The ArchivU project aims to connect the formal evolution of council’s minutes with the transformations of the French University as an institution from the 70s to nowadays. This contribution focuses on a XML-TEI corpus of council‘s minutes of the University of Nanterre enriched with metadata relating to the status and gender of members. Our research aims 1) to identify discursive routines, i.e. semi-fixed segments fulfilling discursive functions (Née et al. 2016), through textual data analysis (cooccurrence, specificity scores, correspondence analysis) and 2) to relate these routines to speaker status over time, using the variables mentioned above.

The speaker status being defined within the University, this study will help understanding the evolution and continuity of the institution itself, *via* the way the institution shapes its voices.

Methodological Exploration of Pattern Mining to Identify Discursive Routines in Formal Records

This communication explores the methodological and epistemological implications of state-of-the-art pattern mining methods in a discourse analysis approach: the inductive extraction of characteristic

discursive routines from corpora of meeting proceedings. The following parameters will be addressed: 1) corpus partitioning (selection of contextual variables, textual segmentation); 2) representation of the relationship between items: sequential patterns (Mekki 2022) or Recurring Lexicosyntactic Trees (Kraif & Diwersy 2012); 3) statistical measure available for contrasting corpora: growth rate (Quiniou et al. 2012) or specificities (Lafon 1981); 4) filtering of initial patterns: minsup (Béchet et al. 2015) or specific vocabulary; 5) data model: morphosyntactic and/or lexical properties. These options represent various ways of constructing the observable features of discursive routines, which we aim to compare based on corpora that vary according to the institutional contexts (university councils or parliamentary assemblies) and language (French, English).

Political Dimension of Councils' Minutes in Small Towns

Local politics has mainly been studied by looking at large cities (Torrekens, 2012; Hijino, 2017). We will focus on a corpus comprising the minutes of 30 small cities councils (less than 10.000 inhabitants) located in France and Switzerland. In small cities, members of councils are non-professional politicians and the minutes are drafted by non-professional writers.

How do external parameters such as the size, location and political affiliation of each city are reflected in these texts? We will examine our corpus combining two data analysis methods: (1) a content analysis method (Reinert, 1983) in order to detect and quantify the themes, linking them to identifiable spheres (political, legal, economic, etc.); (2) a textometric method (based on collocation and specificity score calculation, Heiden *et al.* 2010) in order to identify discursive routines in relation to each sphere.

Religious-Regulated Discourse in 16th-Century Spain: Exploring Inquisitorial Trials during the Reformation

Following the Fifth Lateran Council (1512-1515), the Spanish Inquisition became concerned with new forms of spiritual deviance that it had to define due to the absence of an operative jurisprudential apparatus and theological consensus (Bœglin et al. 2018). Before the Council of Trent (1547-1563), alongside efforts to detect heretical ritual practices, judges discovered and gradually began to regulate Catholic discourse (Dedieu 1978; Betrán Moya et al. 2016), even though religious controversy was acknowledged by the Church as essential to the vitality of dogma. By the late 1550s, however, confessional frontiers had been firmly established, and the repression of Spanish Lutherans became unrelenting (Thomas 2001).

Using two sets of trials from the *Archivo Histórico Nacional* in Madrid—records from 1525-1539, examined by the Toledo Inquisition against the so-called *alumbrados*, and cases from the late 1550s from Valladolid against Protestants—we propose to create two sub-corpora, primarily based on witness statements. To highlight how the processes of regulation took shape, we will contrast various textometric observables that are predominant in one corpus or the other (Lebart et al. 1994; Guilhaumou 2002).

List of references

Bakhtin, M. (1984). Les genres du discours. In: Esthétique de la création verbale. Paris: Gallimard.
Béchet, N., Cellier, P., Charnois, T. & Crémilleux, B. (2015). Sequence mining under multiple constraints. In: Proceedings of the 30th Annual ACM Symposium on Applied Computing, 908-914.
Betrán Moya, J. L., Moreno Martínez, D. & Hernández, B. (2016). Identidades y fronteras culturales en el mundo ibérico en la Edad Moderna.
Bœglin, M., Fernández Terricabras, I. & Kahn, D. (2018). Reforma y disidencia religiosa: la recepción de las doctrinas reformadas en la Península Ibérica en el siglo XVI, Madrid: Casa de Velázquez.
Bourdieu, P. (1982). Langage et pouvoir symbolique. Paris: Seuil.
Dedieu, J.-P. (1978). Les causes de foi de l'Inquisition de Tolède (1483 – 1820). Essai statistique. Mélanges de la Casa de Velázquez, 14, 144-169.
Erjavec, T., Kopp, M., Ljubešić, N. et al. (2024). ParlaMint II: advancing comparable parliamentary corpora across Europe. Language Resources and Evaluation. https://doi.org/10.1007/s10579-024-09798-w
Guilhaumou, J. (2002). Le corpus en analyse de discours : perspective historique. Corpus, 1. https://doi.org/10.4000/corpus.8.
Heiden, S., Magué, J.-P. & Pincemin, B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement. In: Bolasco, S., Chiari, I. & Giuliano, L. (eds) :

Statistical Analysis of Textual Data -Proceedings of 10th International Conference JADT 2010. Roma : Edizioni Universitarie di Lettere Economia Diritto.
Hijino, K. V. (2017). Local Politics and National Policy: Multi-level conflicts in Japan and Beyond. London; Routledge.
Kraif, O. & Diwersy, S. (2012). Le Lexicoscope : un outil pour l'étude de profils combinatoires et l'extraction de constructions lexico-syntaxiques. In : Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, Grenoble, France, 399–406.
Lafon, P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. Mots, 1, 127–165.
Lebart, L. & Salem, A. (1994). Statistique textuelle. Paris : Dunod.
Mekki, J. (2022). Caractérisation de registres de langue par extraction de motifs séquentiels émergents. PhD thesis, Université de Rennes.
Née, É., Sitri, F. & Veniard, M. (2016), Les routines, une catégorie pour l'analyse de discours : le cas des rapports éducatifs. Lidil, 53, 71-93.
Quiniou, S., Cellier, P., Charnois, T. & Legallois, D. (2012). Fouille de données pour la stylistique : cas des motifs séquentiels émergents. In: Actes des 11es Journées Internationales d'Analyse Statistique des Données Textuelles (JADT'12), Liège, Belgique. 821-833.
Reinert, M. (1983). Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte. Les cahiers de l'analyse des données, 187-198.
Thomas, W. (2001), La represión del protestantismo en España, 1517-1648, Leven : Leuven University Press.
Torrekens, C. (2012). Concertation et négociation à l'échelle politique locale. Le cas de la gestion locale de l'islam à Bruxelles. Participations, 2 (1), 126-145. https://doi.org/10.3917/parti.002.0126.

Corpus-based comparative discourse analysis: Methodological challenges and technical affordances

Julia Krasselt[1], **Isabelle Suremann**[1], **Rachelle Vessey**[2], **Niall Curry**[3], **Stephanie Evert**[4], **Philipp Heinrich**[4], **Michael Bender**[5]

[1]Zurich University of Applied Sciences; [2]Carleton University; [3]Manchester Metropolitan University; [4]Friedrich-Alexander-Universität Erlangen-Nürnberg; [5]Technische Universität Darmstadt

Comparison is a core analytical technique in corpus linguistics. From a discourse analytic perspective, it can sharpen our understanding of discourses, support the identification of alternative interpretations, and contribute to the analysis of less explicit phenomena such as discursive absences (Dreesen & Judkowiak, 2011; Storjohann & Schroeter, 2013). It is therefore not surprising that a growing number of researchers are interested in multilingual comparative discourse analysis. However, such analyses pose specific challenges both in terms of methodology and corpus infrastructure (see, among others Vessey, 2013; Dreesen & Czachur, 2019; Taylor & del Fante, 2020; Curry, 2021). For example, we may wonder how equivalence in data and annotation can be established or whether our tertium comparationis is aligned with our research objective (i.e., the underlying dimension or feature that serves as the basis for comparison). It is crucial to pay attention to these questions to ensure that our analysis and interpretation are not influenced predominantly by differences, for example, in language systems themselves or in country-specific media landscapes.

This panel explores the methodological and technical issues of corpus-based comparative discourse analysis with a focus on corpus design and specific methods to bridge linguistic and conceptual differences across multilingual corpora. It is loosely linked to the OSCARS project Making Open Research Data Suitable for Comparative Discourse Analysis (MORCDA). This project aims to enhance data utilisation, facilitate community building and optimise research infrastructures for the needs of comparative research.

Individual Contributions

The panel consists of five presentations followed by a joint discussion. It begins with a talk on the challenges in designing comparable corpora based on discourse models, followed by two contributions addressing different corpus types and annotation strategies. The final two presentations explore computational and linguistic methods for aligning linguistic and conceptual differences. Together, the panel traces a progression from modelling and infrastructure to methodological refinement, highlighting the interaction between technical tools and discourse-analytical aims. The concluding discussion synthesises these perspectives and reflects on best practices for corpus-based comparative discourse analysis.

Examining Swiss-AL: a multilingual corpus infrastructure from Switzerland

Julia Krasselt, **Isabelle Suremann**

After a short introduction on the overall motivation for the panel, the first talk introduces a Swiss corpus infrastructure containing corpora for the four national languages of Switzerland. It serves as an example to illustrate the discourse modelling process in comparative discourse analysis and to investigate the specific challenges of multilingual corpus infrastructures in the modelling process, and how they can be addressed, e.g. questions on data equivalency, and corpora containing both high- and low-resource languages. Drawing from the conclusions on the technical challenges and solutions, the benefits and pitfalls of the multilingual corpus infrastructure for corpus-based comparative discourse analyses across all parts of Switzerland will be discussed.

Examining multilingualism and intercultural dialogue via monolingual data

Rachelle Vessey

The paper addresses how corpus-assisted discourse studies can overcome potential monolingual biases in the examination of multilingualism and intercultural dialogue by using annotation and subcorpora. To demonstrate this, examples of corpus-assisted discourse studies of debates surrounding the United Nations language policy will be shown (McEntee-Atalianis & Vessey, 2020, 2024). It will be demonstrated how annotation and the creation of subcorpora enable researchers to account for diverse intersections within the data, with potential impacts for the study of critical sociolinguistic issues.

Investigating the impact of data and research design on research outcomes

Niall Curry

The paper compares two case studies, employing comparable and parallel corpora to support convergent and divergent contrastive discourse analyses. For the former, an analysis of Brexit discourses in English, French, and Spanish is used to demonstrate cross-cultural differences, through a convergent critical discourse analysis based on a comparison of keywords in each corpus. For the latter, a parallel corpus of academic news blogs posts and their translations is used to conduct a divergent transitivity analysis. Through these studies, the paper draws attention to the affordances of contrastive research for supporting corpus-based contrastive discourses analysis and the need for further engagement with theoretical concepts in contrastive linguistics.

Discoursemes and mutlilingual embeddings as a technical basis for comparative discourse analysis

Stephanie Evert

Comparative discourse analysis across languages and cultures faces the challenge of bringing together widely different framings, conceptualisations and linguistic realisations of discursive patterns. This talk proposes “discoursemes” as a quantitative-qualitative unit of analysis that enables researchers to bridge the gap between different languages and multiple corpus-based analyses of related discourses. The approach is further enhanced through cross-linguistic semantic maps of discoursemes and their constellations, created with the help of multilingual embeddings.

Category development for pragmatic annotation as a linguistic bridge in multilingual and comparative discourse analysis

Michael Bender

This talk addresses the role of category development for pragmalinguistic annotation as a linguistic bridge in multilingual discourse analysis. The focus is on the relationship between function and form in different languages. This involves on the one hand similarities and differences at the level of the annotation of pragmalinguistically graspable practices that are relevant across languages in exemplary discourse domains in German and English. On the other hand, it focuses on the level of the different linguistic surfaces. Criteria such as segmentation, granularity, and depth of interpretation play a decisive role for the formation of annotation guidelines and thus for the inter-annotator agreement between human annotators as well as the operationalisation for automation approaches with a view to machine learning. Referencing practices and uncertainty markers in risk discourses will serve as examples.

Joint discussion: Needs, Demands, and Best Practices in Comparative Discourse Analysis

The last slot is dedicated to an interactive discussion to establish the audiences’ experiences in corpus-based comparative discourse analysis and the needs, wants, demands, and best practices. To that end, the panel organisers will prepare a number of key questions arising from the five presentations.

List of references

Curry, N. (2021). Academic writing and reader engagement: Contrasting questions in English, French and Spanish corpora. Routledge.

Dreesen, P., & Czachur, W. (2019). Vergleichende und kontrastive Diskurslinguistik. Prämissen – Prinzipien – Probleme. In G. Rocco & E. Schafroth (Hrsg.), Methoden der vergleichenden Diskurslinguistik. Germanistisch-romanistische Beiträge zur Methodenreflexion und Forschungspraxis. (S. 59–91).

Dreesen, P., & Judkowiak, J. (2011). Passiv im Osten, kollektiv schuldig und selbstverständlich in Europa: Kritik an deutschen und polnischen Schulbüchern des Faches Geschichte mittels kontrastiver Diskurslinguistik. Hempen.

McEntee-Atalianis, L., & Vessey, R. (2020). Mapping the language ideologies of organisational members: A Corpus Linguistic Investigation of the United Nations’ General Debates (1970-2016). Language Policy, 19, 549–573.

McEntee-Atalianis, L., & Vessey, R. (2024). Using corpus linguistics to investigate agency and benign neglect in organisational language policy and planning: The United Nations as a case study. Journal of Multilingual & Multicultural Development, 45(2), Article 2.

Storjohann, P., & Schroeter, M. (2013). Präsenz und Absenz lokaler Diskursgebrauchsmuster am Beispiel des deutschen und britischen Krisendiskurses. In M. Wengeler & A. Ziem (Hrsg.), Sprachliche Konstruktionen von Krisen. Interdisziplinäre Perspektiveen auf ein fortwährend aktuelles Phänomen. (S. 185–208). Hempen. https://centaur.reading.ac.uk/31939/

Taylor, C., & del Fante, D. (2020). Comparing across languages in corpus and discourse analysis: Some issues and approaches. Meta (Montréal), 65(1), 29–50. https://doi.org/10.7202/1073635ar

Vessey, R. (2013). Challenges in cross-linguistic corpus-assisted discourse studies. Corpora, 8(1), 1–26. https://doi.org/10.3366/cor.2013.0032

THURSDAY 3rd JULY

Corpus-assisted CDA approaches to polycrisis

Cinzia Bevitori[1], Niall Curry[2], Dario Del Fante[3], Stefania Maci[4], Rakan Alibri[5], Katherine Elizabeth Russo[6], Virginia Zorzi[7]

[1]University of Bologna; [2]Manchester Metropolitan University; [3]University of Ferrara; [4]University of Bergamo; [5]Tabuk University; [6]University of Naples L'Orientale; [7]University of Turin

The noun polycrisis is increasingly used by organisations such as UNICEF and the WEF to encapsulate the intersecting, accelerating challenges facing the world today. Krzyżanowski et al. (2023, 423) define polycrisis as the 'combination of many, more or less simultaneous and overlapping, crises whose repercussions unfold in a cumulative manner'. Polycrisis occurs 'when crises in multiple global systems become causally entangled in ways that significantly degrade humanity's prospects' (Lawrence et al. 2022, 2).

According to Kluth (2023), the world has seen intersecting crises before, so neologisms like polycrisis are unnecessary. Similarly, Henig and Knight (2023) call for research to explain what exactly makes polycrisis characteristic of the present epoch. Existing research offers some basis for this temporal focus and a rebuttal of Kluth's argument. Krzyżanowski et al. (2023, 416) state that connections between crises have become more intense since the 'turn of the new millennium'. Lawrence et al. (2024, 5) argue that the world is 'far more connected' than during previous intersecting crises. This context suggests a pressing need to address polycrisis now - a call we take up in this thematic panel.

We embrace polycrisis as the overarching theme for this panel because we agree with Lawrence et al. (2024, 2) that 'the polycrisis concept – if defined clearly and translated into a productive program of research and action' – can help us to better understand the society we live in. Within the 3-hour thematic panel, we address the intersecting crises of risks to life, climate change, homelessness and the cost of living, and migration using corpus-assisted critical discourse analytical methods. We consider each of the aforementioned factors separately while also considering how the crises interact and exacerbate one another to form a conjoined polycrisis.

The panel begins with a presentation by Rakan Alibri, titled 'From Risks to Polycrsis: A Corpus-Assisted Critical Discourse Analysis of Media Constructions of Risks to Life'. News media have a critical role in influencing risk perception, and this paper presents a corpus-assisted critical discourse analysis approach used to identify discursive strategies employed in the construction of risks to life in the British press. The author illustrates how these strategies can be identified, and explores how they may amplify or attenuate risk perception. In conclusion, the need for comparing risks to each other and broadening the analysis of the risk context in order to identify any potential polycrisis in society is emphasized.

The next session is a paper by Cinzia Bevitori and Katherine Russo, titled '"Because climate change is the crisis that will stay with us": Crisis, Polycrisis, Permacrisis in the EU discursive space'. The paper focuses on a purpose-built corpus of policy communications addressing health and climate change within the EU's discursive space. Using a corpus-assisted approach to framing analysis, the paper serves two main purposes. First, it argues that the concept of "polycrisis" functions as a rhetorical tool to legitimize specific actions, thereby reinforcing the EU's collective identity. Second, it reflects on the methodological benefits and constraints of using corpus-assisted techniques for qualitative coding and interpretation of framing.

The third paper session is by Stefania Maci, titled 'From climate change to global crises. The perspective of UNO'. To encourage action on climate change, the UNO Agenda 2030 redirects to the ACT NOW campaign site and the Climate Change website. Drawing on corpus linguistics, CDA, and the Discourse-Historical Approach (DHA), this presentation analyzes if, and how, crises deriving from climate change are discursively communicated in official 2023 UNO documents. Maci concentrates on the linguistic analysis of three key terms (crisis and crises, challenge(s), and burden(s)). Results indicate that responsibility for the processes in UNO documents is never specified. Instances of personification, passivization, and nominalization contribute to making the texts more vague in terms of social actors' responsibility, thus hiding agency and relegating responsibility for actions to the background.

The fourth paper, by Niall Curry and Gavin Brookes, is titled "The discursive framing of the climate and health polycrisis in English, French and Spanish". This paper investigates the complexities underpinning this polycrisis through an analysis of its discursive framing, through a corpus-based contrastive analysis

of the use of health, santé and salud in climate-themed and health-themed parascientific communication in English, French, and Spanish. Three recurrent framing activities are identified: these include defining and contextualising the polycrisis, representing cause and effect, and proposing solutions. The findings highlight the crucial role of cultural and linguistic diversity in shaping responses to global crises, and call for pluriversal approaches to knowledge production to address the complex challenges posed by global polycrises.

The concluding paper, by Dario Del Fante and Virginia Zorzi, is titled "Polycrisis and its relevance to migration discourses. A corpus-assisted investigation". The authors investigate the potential relevance of the polycrisis concept in the discursive construction of migration in news corpus focused on the topic of migration and in a corpus of personal narratives by UK-based refugees. They explore the patterns of use of lemmas including "crisis", "migration", "migrant", "refugee", "asylum seeker", synonyms of "crisis", and they conduct an analysis of the semantic fields potentially related to the representation of migration as a crisis. Several dimensions emerge in the news corpus (e.g., migration as a crisis in itself, as a phenomenon listed together with other negative global trends) and in the narrative corpus (e.g., migration as a personal crisis, migration as caused by a situation of crisis).

List of references

Estes, Carroll L. (1983). "Social Security: The Social Construction of a Crisis." The Milbank Memorial Fund Quarterly. Health and Society, 61(3): 445–61.

Kluth, A., 2023. So we're in a polycrisis. Is that even a thing. The Washington Post.

Krzyżanowski, Michał, Ruth Wodak, Hannah Bradby, Mattias Gardell, Aristotle Kallis, Natalia Krzyżanowska, Cas Mudde, and Jens Rydgren. (2023). Discourses and practices of the 'New Normal'. Towards an interdisciplinary research agenda on crisis and the normalization of anti- and post-democratic action. Journal of Language and Politics, 22(4): 415-437.

Lawrence, Michael, Scott Janzwood, and Thomas Homer-Dixon. (2022). What is a global polycrisis? Technical paper.

Lawrence, Michael, Thomas Homer-Dixon, Scott Janzwood, Johan Rockstöm, Ortwin Renn, and Jonathan F. Donges. (2024). Global polycrisis: the causal mechanisms of crisis entanglement. Global Sustainability, 7: 1-14.

Parnell, Tamsin & Van Hout, Tom & Del Fante, Dario. (2025). Critical discursive responses to Polycrisis. In: Parnell, T., Van Hout, T., and Del Fante, D. (Eds.) Critical Approaches to Polycrisis. Discourses of Conflict, Migration, Risk and Climate. London: Palgrave.

PAPER PRESENTATIONS

How can corpus tools help in understanding political stance? A corpus-assisted study on parliamentary debates

Rahma Al-Busafi

University of Technology & Applied Sciences

This study explores the use of corpus tools to analyze and understand stance in political discourse, with a particular focus on the subgenre of parliamentary debates. Specifically, it investigates how stance is linguistically realized in the UK Parliament, with emphasis on debates held in the House of Commons. Parliamentary discourse, characterized by its inherently opinionated nature and the interaction of MPs representing diverse political affiliations, offers a rich context for studying stance as a pervasive and dynamic phenomenon.

To achieve this, a corpus of parliamentary debates from the 2010–2015 parliamentary cycle on the issues of 'flooding' has been compiled. Stance is analyzed by identifying appraisal resources and evaluative language using the Appraisal Framework (Martin and White, 2005). The UAM Corpus Tool is employed to annotate debates for Appraisal resources based on MPs' parliamentary roles, while AntConc software is used to identify stance-related patterns and evaluate appraisal moves in the discourse. Additionally, the Hansard corpus (1803–2005) is consulted to contextualize these findings and uncover recurring patterns in stance-taking strategies. Key instances from the corpus are presented to highlight the most prominent linguistic features of stance in parliamentary debates.

The findings reveal that the relationship between language and political stance is complex and highly nuanced. The study argues that political stance cannot always be directly inferred from surface-level language forms. Politicians often adopt public stances that may differ significantly from their private positions, driven by the need to navigate diplomatic and political objectives.

This research underscores the value of corpus-assisted methods in uncovering the intricate dynamics of political stance in parliamentary debates. Bridging corpus tools with discourse analysis, it highlights the nuanced interplay between language, political ideologies, and rhetorical strategies, contributing to a deeper understanding of political communication and its role in democratic processes.

List of references

Martin, J., and White, P. (2005). The Language of Evaluation. Basingstoke: Palgrave Macmillan.

A comprehensive critical review of NLP and corpus linguistics in Pharmacovigilance

Doaa Al-Turkey, Rasheed Mohammad, Andrew Wilson, Tatiana Grieshofer

Birmingham City University

Pharmacovigilance- the science of detecting, assessing, and preventing adverse drug reactions (ADRs)-is essential for ensuring drug safety and protecting public health. In an era where healthcare data is increasingly unstructured and originates from sources such as electronic health records, clinical notes, and social media, traditional monitoring methods are challenged. This review draws on insights from over two hundred studies to explore how Natural Language Processing (NLP)- computer-based methods that analyse human language- and Corpus Linguistics can enhance pharmacovigilance-systematic techniques, such as n-gram analysis, keyword frequency lists, collocation analysis, and semantic annotation—that reveal patterns and language trends critical for identifying ADRs.

These methods offer improved scalability, real-time monitoring, and the ability to process data in multiple languages. However, challenges remain variability in data quality, complex domain-specific language, linguistic ambiguity, and ethical concerns around using patient-generated data. Moreover, the lack of standardised benchmarks for algorithm validation and limited interdisciplinary collaboration between computational linguists and pharmacovigilance experts pose additional hurdles.

Given the critical role of pharmacovigilance in mitigating drug risks and protecting patient health, understanding and advancing these computational techniques is vital. Emerging trends such as transformer-based language models (BERT, GPT) and hybrid human-AI workflows-are capable of transforming pharmacovigilance. This review calls for the development of rigorous, transparent, and ethically responsible frameworks to fully harness the potential of NLP and corpus linguistics, ultimately contributing to safer, evidence-based global drug practices.

Variation in phrase frame structure and function in argumentative writing by Saudi and native English-speaking students

Basim Alamri, Assem Alqarni

King Abdualaziz University

Recent research has increasingly focused on discontinuous formulaic sequences, particularly phrase frames (p-frames), which are key to understanding language use in writing. Investigating p-frames in English as a Foreign Language (EFL) writing can enhance insights into second language (L2) writing and teaching practices. Utilizing a corpus-driven approach, this study aims to identify and compare the p-frames used by Saudi EFL university students in argumentative essays with those employed by native English speakers. The study involved compiling two comparable learner corpora: a Saudi corpus and LOCNESS (Granger, Dagneaux, Meunier, Paquot, 2009). The Saudi corpus included 500 argumentative essays (about 165,140 words) of students' final writing exams, while the LOCNESS consisted of 175 argumentative essays (148,516 words). The Saudi students' level was considered as B1 according to the Common European Framework of Reference for Languages (CEFR). Several procedures were undertaken to analyze data from the corpora. The study selected the 100 most frequent p-frames using the AntConc 4.1.4 corpus analysis software, which were then manually filtered. The p-frames used by two groups were categorized based on their frequency, structure, and function. The findings revealed that EFL writers used significantly fewer p-frames than native speakers. Moreover, EFL writers demonstrated a greater reliance on more predictable p-frames. In terms of functionality, EFL writers were more inclined to use stance and discourse-organizing p-frames compared to their native-speaking counterparts. This suggests that EFL writers tend to favor expressions they find more familiar and comfortable with, and unlike native speakers, they more frequently utilize highly predictable expressions in argumentative essays. The study has pedagogical implications for the globalized educational environment and contributes to a more complex understanding of phraseological proficiency among EFL learners.

List of references

Granger, S., Dagneaux, E., Meunier, F., & Paquot, M. (Eds.). (2009). International corpus of learner English. Louvain-la-Neuve: Presses universitaires de Louvain.

Leveraging AI for translation education: A corpus-based framework using SauLTC and LLMs

Moneerh Aleedy[1], Maha Al-Harthi[1], Fatma Alshihri[1], Souham Meshoul[1], Salwa Alramlawi[1], Badr Aldaihani[1], Hadil Shaiba[1], Eric Atwell[2]

[1]Princess Nourah bint Abdulrahman University; [2]University of Leeds

Translation education (TE) requires innovative solutions to address its labor-intensive nature and the lack of tailored bilingual corpora, particularly for English-Arabic translation. This study introduces a novel framework utilizing the Saudi Learner Translation Corpus (SauLTC) and large language models (LLMs) to develop high-quality parallel sentence datasets designed for TE. The research focuses on transforming SauLTC into a didactic resource by employing GPT models for sentence alignment, supported by multilingual embedding techniques like LaBSE and MPNet. Data preparation involved cleaning and aligning English-Arabic sentence pairs while addressing inconsistencies in grammar, context, and cultural variations. The alignment process combined cosine similarity metrics with semantic analysis, achieving an 85.2% alignment accuracy using LaBSE in conjunction with GPT. Human evaluation, based on Multidimensional Quality Metrics (MQM), further validated the dataset with an exceptional 98% quality score. The resulting dataset comprises 15,845 parallel sentences, 95% of them are above a 0.7 similarity threshold, annotated to highlight structural and lexical equivalences. These annotations facilitate student engagement with critical aspects of translation, such as error analysis, problem-solving, and post-editing strategies. The study demonstrates how AI can bridge gaps in TE, providing personalized tools that enhance learning experiences while fostering autonomy in translation practices. The implications extend to improving translator training by aligning educational practices with industry demands for AI-augmented workflows. This framework sets a precedent for integrating corpus linguistics and AI-driven methodologies, contributing to cross-linguistic research and advancing the global discourse on translation education.

List of references

Al-Batineh M, Al Tenaijy M. 2024. Adapting to technological change: An investigation of translator training and the translation market in the Arab world. Heliyon 10. DOI: 10.1016/j.heliyon.2024.e28535.

Alfarizy G, Mandala R. 2022. Verification of Unanswerable Questions in the Question Answering System using Sentence-BERT and Cosine Similarity. 2022 9th International Conference on Advanced Informatics: Concepts, Theory and Applications, ICAICTA 2022. DOI: 10.1109/ICAICTA56449.2022.9932903.

Al-Harthi M, Al-Saif A. 2019. The Design of the SauLTC application for the English-Arabic Learner Translation Corpus. In: Proceedings of the 3rd Workshop on Arabic Corpus Linguistics. Association for Computational Linguistics, 80–88.

Al-Harthi M, Alsaif A, Al-Nafjan E, Alshihri F, Saleh M. 2024. Saudi Learner Translation Corpus: The design and compilation of an English–Arabic learner translation corpus. PLOS ONE 19:1–22. DOI: 10.1371/journal.pone.0303729.

Bisiada M. 2017. Universals of editing and translation. In: Empirical modelling of translation and interpreting. Language Science Press, 241–275. DOI: 10.5281/ZENODO.1090972.

Brien SO". 2012. Towards a Dynamic Quality Evaluation Model for Translation. Journal of Specialised Translation.

Castagnoli S. 2020. Translation choices compared: Investigating variation in a learner translation corpus. In: Granger S, Lefer M-A eds. Translating and Comparing Languages: Corpus-based Insights. 25–44.

Chimoto EA, Bassett BA. 2022. Very Low Resource Sentence Alignment: Luhya and Swahili.

Freitag M, Foster G, Grangier D, Ratnakar V, Tan Q, Macherey W. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. Transactions of the Association for Computational Linguistics 9:1460–1474. DOI: 10.1162/TACL_A_00437/108866/EXPERTS-ERRORS-AND-CONTEXT-A-LARGE-SCALE-STUDY-OF.

Garbovskiy N, Kostikova O. 2020. Translation Didactics: What Are the Ways to Train a Translation Teacher? New Frontiers in Translation Studies:45–61. DOI: 10.1007/978-981-15-7390-3_4.

Kanglang L, Afzaal M. 2021. Artificial intelligence (AI) and translation teaching : a critical perspective on the transformation of education. International journal of educational sciences 33:64–73. DOI: 10.31901/24566322.2021/33.1-3.1159.

Kenny D. 2019. Technology and translator training. In: The Routledge Handbook of Translation and Technology. Routledge, 498–515. DOI: 10.4324/9781315311258-35.
Kong L. 2022. [Retracted] Artificial Intelligence-Based Translation Technology in Translation Teaching. Computational Intelligence and Neuroscience 2022:6016752. DOI: 10.1155/2022/6016752.
Kübler N, Mestivier A, Pecman M. 2022. Using comparable corpora for translating and post-editing complex noun phrases in specialized texts. In: Lefer SG& M-A ed. Extending the Scope of Corpus-Based translation Studies. Bloomsbury Advances in Translation. Bloomsbury Publishing, 237–266.
Kurek J, Latkowski T, Bukowski M, Świderski B, Łępicki M, Baranik G, Nowak B, Zakowicz R, Dobrakowski Ł. 2024. Zero-Shot Recommendation AI Models for Efficient Job–Candidate Matching in Recruitment Process. Applied Sciences 2024, Vol. 14, Page 2601 14:2601. DOI: 10.3390/APP14062601.
Lapshinova-Koltunski E. 2022. Detecting normalization and shining-through in novice and professional translations. In: Granger S, Lefer M-A eds. Extending the Scope of Corpus-Based Translation Studies. London: Bloomsbury Academic, 182–206. DOI: 10.5040/9781350143289.0015.
Lefer M-A. 2020. Parallel Corpora. In: Paquot Magali and Gries STh ed. A Practical Handbook of Corpus Linguistics. Cham: Springer International Publishing, 257–282. DOI: 10.1007/978-3-030-46216-1_12.
LEFER M-A, PIETTE J, BODART R. 2022. Machine Translation Post-Editing Annotation System (MTPEAS) manual.
Lommel A, Gladkoff S, Melby A, Wright SE, Strandvik I, Gasova K, Vaasa A, Benzo A, Sparano RM, Foresi M, Innis J, Han L, Nenadic G. 2024. The Multi-Range Theory of Translation Quality Measurement: MQM scoring models and Statistical Quality Control.
Mohamed YA, Khanan A, Bashir M, Mohamed AHHM, Adiel MAE, Elsadig MA. 2024. The Impact of Artificial Intelligence on Language Translation: A Review. IEEE Access 12:25553–25579. DOI: 10.1109/ACCESS.2024.3366802.
Mohsen MA. 2024. Artificial Intelligence in Academic Translation: A Comparative Study of Large Language Models and Google Translate. PSYCHOLINGUISTICS 35:134–156. DOI: 10.31470/2309-1797-2024-35-2-134-156.
Park D, Padó S. 2024. Multi-Dimensional Machine Translation Evaluation: Model Evaluation and Resource for Korean. :11723–11744.
Rodríguez De Céspedes B. 2019. Translator Education at a Crossroads:the Impact of Automation. Lebende Sprachen 64:103–121. DOI: 10.1515/LES-2019-0005/MACHINEREADABLECITATION/RIS.
Saldías B, Foster G, Freitag M, Tan Q. 2022. Toward More Effective Human Evaluation for Machine Translation. HumEval 2022 - 2nd Workshop on Human Evaluation of NLP Systems, Proceedings of the Workshop:76–89. DOI: 10.18653/V1/2022.HUMEVAL-1.7.
Steigerwald E, Ramírez-Castañeda V, Brandt DYC, Báldi A, Shapiro JT, Bowker L, Tarvin RD. 2022. Overcoming Language Barriers in Academia: Machine Translation Tools and a Vision for a Multilingual Future. Bioscience 72:988. DOI: 10.1093/BIOSCI/BIAC062.
Uzar R, Waliński JT. 2001. Analysing the Fluency of Translators. International Journal of Corpus Linguistics 6:155–166. DOI: 10.1075/IJCL.6.SI.12UZA/CITE/REFWORKS.
Wang Y. 2023. Artificial Intelligence Technologies in College English Translation Teaching. Journal of Psycholinguistic Research 52:1525–1544. DOI: 10.1007/S10936-023-09960-5/TABLES/4.
Wu T, He S, Liu J, Sun S, Liu K, Han QL, Tang Y. 2023. A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development. IEEE/CAA Journal of Automatica Sinica 10:1122–1136. DOI: 10.1109/JAS.2023.123618.
Wurm A. 2020. Translation quality in an error-annotated translation learner corpus. In: Granger S, Lefer Marie-Aude eds. Translating and Comparing Languages: Corpus-based Insights.. Presses universitaires de Louvain,.
Yang R, Takechi K, Zhang Y, He Y. 2021. Challenges and Countermeasures of Translation Teaching in the Era of Artificial Intelligence. Journal of Physics: Conference Series 1881:022086. DOI: 10.1088/1742-6596/1881/2/022086.
Zaghlool ZDM, Khasawneh MAS. 2024. Aligning Translation Curricula with Technological Advancements; Insights from Artificial Intelligence Researchers and Language Educators. Studies in Media and Communication 12:58–70.
Zhao J, Li D, Tian L. 2020. Translation Education. Springer Singapore. DOI: https://doi.org/10.1007/978-981-15-7390-3.

The use of English within Arabic by Saudi YouTubers

Huda Alkhamali

University of Southampton; University of Hail

The use of English words in Saudi Arabic has risen recently, especially in digital communication (Mahboob & Elyas, 2014). This trend aligns with Saudi Vision 2030, which emphasizes English proficiency as vital for economic diversification and global integration (Saudi.Gov, 2016). The increased integration of English into Saudi discourse reflects both societal change and strategic national goals.

This study explores this phenomenon using the ESATube Corpus, a specialized collection of 570,169 words from 70 hours of video content by prominent Saudi YouTubers (January 2023–December 2024). It includes the following genres: gaming, food vlogging, entertainment, book reviews, and technology reviews. By analyzing informal usage of English within Saudi Arabic across diverse YouTube genres, the study investigates patterns that reflect the sociolinguistic impact of Saudi vision 2030. The aim is to investigate the use, frequency, and context of these words by analyzing concordance lines and collocates to address these questions: What patterns emerge in English usage among Saudi YouTubers? How does this usage vary across genres?

Preliminary analysis classifies English words into three categories. First are established loanwords like video, camera, and telephone, fully integrated into Arabic, accepted by the speech community, and thus achieving loanword status (Poplack et al., 1988). Second are code-switching instances, often found in gaming and book review content. Third are newer borrowings such as wow, like, and nice, which include nouns, verbs, adjectives, and interjections. Unlike older loanwords, which were mainly nouns filling lexical gaps, these often have Arabic equivalents but spread quickly through digital communication.

Initial findings suggest that this hybrid style reflects both the impact of English as a global language and the adaptability of Arabic highlighting language contact linked to Saudi's national goals. This study contributes to language contact research in digital communication by focusing on the relatively underexplored Saudi Arabic-English interaction in spoken digital discourse.

List of references

Mahboob, A., & Elyas, T. (2014). English in the kingdom of Saudi Arabia. World Englishes, 33(1), 128-142.

Poplack, S.,Sankoff, D., & Miller, C. R. (1988). The social correlates and linguistic processes of lexical borrowing and assimilation. Linguistics, 26(1), 47-104.

Saudi.Gov. (2016). Vision 2030. Retrieved 29 May from https://www.vision2030.gov.sa/v2030/overview/

Data-driven learning in context: A novel meta-analysis approach

Reem Alojaimi[1,2]

[1]Lancaster University; [2]King Saud University

Data-driven learning (DDL) is an approach to language teaching providing corpora access to language learners (Johns, 1991). Its effectiveness in (second) language acquisition has been demonstrated in literature (Boulton & Vyatkina, 2021; O'Keeffe, 2021; Pérez-Paredes, 2022) and synthesised by recent meta-analyses (Boulton & Cobb, 2017; Cobb & Boulton, 2015; Lee et al., 2019; Mizumoto & Chujo, 2015; Ueno & Takeuchi, 2023; Yoon & Lee, 2024), attracting constant research attention. This study expands on previous syntheses by investigating the effect of learner-related factors (population variables) and intervention-related factors (treatment variables) in the classroom context and their influence on effect size. It employs two methods: a simple meta-analysis and a multiple meta-regression analysis using a multi-model approach. 97 studies met the inclusion criteria, and 125 unique between-subjects samples with over 6,000 participants were included.

The results show that DDL is an effective teaching approach with a large effect (g = 0.95). Among the population variables, region and sample size predict a larger effect size. On the other hand, institution type and students' language proficiency are not important predictors. Among the treatment variables, corpus type, linguistic target of intervention and DDL interaction type (i.e. direct vs indirect DDL) are, in order, the most important predictors. Although experiment duration is not considered an important predictor of DDL's effectiveness, further analysis shows that longer interventions could lead to more substantial learning outcomes. Further analysis also revealed the versatility of DDL: despite its application being largely exclusive to research and universities (Timmis &Templeton, 2023, p.420), the findings prove that it is also promising in high schools as well. Overall, the meta-analysis suggests that medium- and long-term interventions with direct access to public corpora in medium-sized groups of learners are the most effective conditions in DDL for better learning gains. Implications for future studies will be discussed.

List of references

Boulton, A., & Cobb, T. (2017). Corpus Use in Language Learning: A Meta-Analysis. Language Learning, 67(2), 348–393. https://doi.org/10.1111/lang.12224

Boulton, A., & Vyatkina, N. (2021). Thirty years of data-driven learning: Taking stock and charting new directions over time. http://hdl.handle.net/10125/73450

Cobb, T., & Boulton, A. (2015). Classroom applications of corpus analysis. In D. B. & R. Reppen (Ed.), Cambridge Handbook of Corpus Linguistics (pp. 478–497). Cambridge University Press. https://doi.org/10.1017/CBO9781139764377.027

Johns, T. (1991). "Should you be persuaded": Two samples of data-driven learning materials. ELR Journal, 4, 1–16.

Lee, H., Warschauer, M., & Lee, J. H. (2019). The Effects of Corpus Use on Second Language Vocabulary Learning: A Multilevel Meta-analysis. Applied Linguistics, 40(5), 721–753. https://doi.org/10.1093/applin/amy012

Mizumoto A., & Chujo K. (2015). A Meta-analysis of Data-driven Learning Approach in the Japanese EFL Classroom. English Corpus Studies=英語コーパス研究, 22, 1–18.

O'Keeffe, A. (2021). Data-driven learning – a call for a broader research gaze. Language Teaching, 54(2), 259–272. https://doi.org/10.1017/S0261444820000245

Pérez-Paredes, P. (2022). A systematic review of the uses and spread of corpora and data-driven learning in CALL research during 2011–2015. Computer Assisted Language Learning, 35(1–2), 36–61. https://doi.org/10.1080/09588221.2019.1667832

Ueno, S., & Takeuchi, O. (2023). Effective corpus use in second language learning: A meta-analytic approach. Applied Corpus Linguistics, 3(3), 100076. https://doi.org/10.1016/j.acorp.2023.100076

Yoon, K.-H., & Lee, D. J. (2024). A Meta-Analysis of Data-Driven Learning (DDL) in EFL/ESL Settings. Brain, Digital, & Learning, 14(2), 283–304. https://doi.org/10.31216/BDL.20240017

Lexical bundles and speech fluency in British and international university students' talk

Federica Barbieri[1], Yusuf Ozturk[2]

[1]Swansea University; [2]Mus Alparslan University

Formulaic language has been studied under many rubrics and from different theoretical and methodological perspectives. One of these involves using frequency-based criteria to identify multi-word sequences. Lexical bundles are sequences of three or more words that frequently recur in a corpus, in texts representing a range of speakers/writers (Biber et al., 2004). Research shows that they are associated with speech fluency (Tavakoli & Uchihara, 2020) and oral proficiency (Kyle & Crossley, 2015): a large repertoire of lexical bundles reduces cognitive load (Tremblay et al., 2011) and enhances fluency (Yan, 2019). However, few studies have investigated the relationship between use of formulaic sequences like lexical bundles and fluency (e.g. Tavakoli & Uchihara, 2020). Further, no studies have distinguished between structurally complete and incomplete bundles in relation to speech fluency. Yet structurally complete sequences are processed differently and are likely to have a greater processing advantage (Jeong & Jiang, 2019).

Accordingly, the present study examines the relationship between lexical bundles use and speech fluency in the casual talk (in English) of British and international students at a UK university. The study is based on lexical bundles occurring in two corpora of university student talk, representing British students' and international students' talk, and amounting to 140,000 and 87,000 words respectively. The corpora comprise semi-structured interviews with university students at a British university. For the analysis of fluency, an automated script (De Jong & Wempe, 2008) was used in PRAAT to calculate speaking and pause times. Three- and four-word bundles were extracted using Wordsmith Tools 6 (Scott, 2011). Statistical analyses examined the relationship between lexical bundles, and speech rate, articulation rate, and number and length of pauses. Preliminary findings suggest that greater and more varied use of lexical bundles is associated with higher articulation rate and with fewer and shorter pauses.

List of references

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at …: Lexical Bundles in University Teaching and Textbooks. Applied Linguistics, 25(3). https://doi.org/10.1093/applin/25.3.371

De Jong, N., & Wempe, T. (2008). Praat Script Syllable Nuclei. Retrieved 7/11/2024 from https://sites.google.com/site/speechrate/speech-rate-praat-script-that-detects-syllable-nuclei/praat-script-syllable-nuclei

Jeong, H., & Jiang, N. (2019). Representation and processing of lexical bundles: Evidence from word monitoring. System, 80, 188-198.

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

Scott, M. (2011). Wordsmith Tools (Version 6). Lexical Analysis Software.

Tavakoli, P., & Uchihara, T. (2020). To What Extent Are Multiword Sequences Associated With Oral Fluency? Language Learning, 70(2), 506-547. https://doi.org/10.1111/lang.12384

Tremblay, A., Derwing, B., Libben, G., & Westbury, C. (2011). Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. Language Learning, 61(2), 569-613.

Yan, X. (2019). Unpacking the Relationship Between Formulaic Sequences and Speech Fluency on Elicited Imitation Tasks: Proficiency Level, Sentence Length, and Fluency Dimensions. TESOL Quarterly, 54(2), 460-487. https://doi.org/10.1002/tesq.556

A corpus analysis of idiolectal n-grams

Sadie Barlow, Andrea Nini

University of Manchester

This study explores linguistic individuality - each individual's unique repertoire of units (sequences of words, morphemes or parts of speech) that they use recurrently - through a corpus-based analysis. Whilst previous research tends to focus on collective linguistic features, this study targets fine-grained, individual-specific patterns that can be identified through computational authorship verification techniques. Some of these sequences are highly specific to an individual, such as Tony Blair's use of *entirely understand* (Mollin 2009). However, there is also intuitively overlap across the repertoires of units that different individuals possess, for instance very common lexical bundles such as *I said to him* (Biber et al. 2021). As such, it is more often the combination of a large number of core grammatical constructions, as opposed to a small number of noticeably idiosyncratic phrases, that results in greater variation between authors than within one individual's language (Barlow 2013).

The present study found that across 18 different authors, each writing two summaries of the exact same text 30 days apart, only one character 7-gram featured across all of the texts. All authors used at least one long character n-gram (7-9 characters) in both texts that was entirely unique to them. The study also explores whether the component units within the n-grams differ between what is entirely individual, yet consistent, and what is used consistently, but is shared by other members of the group. The implications of this research centre on enhancing our understanding of why authorship analysis methods work, producing empirical evidence of cognitive linguistic theories of individuality, which a limited number of existing studies have aimed to investigate, and exemplifies the benefits and possibilities of applying corpus linguistic methodologies to authorship analysis problems.

List of references

Barlow, Michael. 2013. Individual Differences and Usage-based Grammar. International Journal of Corpus Linguistics 1, 443-478.

Biber, Douglas, Stig Johansson, Geoffrey N Leech, Susan Conrad & Edward Finegan. 2021. Lexical Expressions in Speech and Writing. In Grammar of Spoken and Written English, 979 – 1030. Amsterdam: John Benjamin's Publishing Company.

Mollin, Sandra. 2009. "I entirely understand" is a Blairism: The Methodology of Identifying Idiolectal Collocations. International Journal of Corpus Linguistics 14, 367-392.

'Up' words are good words. Tracing the association between space and valence across the lexicon using distributional semantics

Sara Bartl, Bodo Winter, Jeannette Littlemore

University of Birmingham

In Western cultures, the concept of 'good' is generally associated with 'up' as opposed to 'down' (Lakoff & Johnson, 1980). This association between space and emotional valence is reflected in multiple facets of Western culture, such as linguistic metaphors (*This is a very uplifting film*, *I'm feeling down*), gesture (thumbs-up, thumbs-down), and cultural artifacts, such as depictions of heaven (up) and hell (down). Additionally, there is strong experimental evidence for the conceptual reality of this mapping (e.g. Casasanto and Dijkstra, 2010; Woodin and Winter, 2018).

In this paper, we examine how this space-valence association shapes the lexicon through distributional semantics. Distributional semantic models are vector representations of word meaning learned from co-occurrence patterns in large corpora (Lenci, 2018). In the distributional space, words that occur in similar contexts are closer together. This spatial representation of meaning affords the use of new methods to interrogate co-occurrence patterns, such as semantic projection. Semantic projection works by creating meaningful dimensions in the word embedding space and assessing how words pattern along it (Grand et al, 2022). For example, to examine a word's valence, we can establish a dimension from "good" to "bad" and then project the vector of a word onto the valence dimension.

Using semantic projection, we show that the space-valence association governs the structure of large parts of the lexicon. We demonstrate that to the extent that words are biased towards an 'up' orientation (e.g., *sun, rainbow, peak*), they are also rated as being more positive, whereas 'down' words (e.g., *bury*, *grave, cockroach*) are rated as being more negative. Additionally, we show that spatial and valence information relating to individual lexical items as obtained from co-occurrence patterns is congruent with ratings provided by human participants, thereby demonstrating close correspondence between meaning patterns derived from corpora and context-free intuitions (cf. Winter, 2022).

List of references

Casasanto, D., & Dijkstra, K. (2010). Motor action and emotional memory. Cognition, 115(1), 179-185.

Grand, G., Blank, I. A., Pereira, F., & Fedorenko, E. (2022). Semantic projection recovers rich human knowledge of multiple object features from word embeddings. Nature human behaviour, 6(7), 975-987.

Lakoff, G., & Johnson, M. (2008). Metaphors we live by. University of Chicago press.

Lenci, A. (2018). Distributional models of word meaning. Annual review of Linguistics, 4(1), 151-171.

Winter, Bodo. (2022) Managing semantic norms for cognitive linguistics, corpus linguistics, and lexicon studies. In Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller & Lauren B. Collister (eds.), The open handbook of linguistic data management. Cambridge, MA: MIT Press.

Woodin, G., & Winter, B. (2018). Placing abstract concepts in space: Quantity, time and emotional valence. Frontiers in Psychology, 9, 2169.

From keywords to keyMWEs: A multiword expressions augmented keyness analysis

Chadi Ben Youssef[1], **Stefan Th. Gries**[2]

[1]University of Neuchâtel; [2]University of California

Keywords or keyness analysis play a central role in determining the linguistic characteristics of a text or corpus compared to a reference corpus. For example, it can reveal how political orientations influence the linguistic features used in speeches and electoral debates (Baker & McEnery, 2005; Egbert & Biber, 2023), determine the different lexical characteristics and discourse strategies deployed in media (Almaghlouth, 2022; Moreno-Ortiz & García-Gámez, 2023), or analyze text genres and varieties (Leech & Fallon, 1992; Scott & Tribble, 2006; Gries, 2021). However, despite the widespread use of keyness analysis and its numerous applications, nearly all studies rely on corpora tokenized on the single-word level.

In this study, we compare a keyness analysis of BNC_{acad} against BNC_{rest} when the corpus is tokenized on the single-word level to the same kind of analysis when the corpus is tokenized in a way that recognizes Multiword Expressions (MWEs) beyond the small number of <mw> types annotated in the corpus. To this end, the BNC was first re-tokenized using the MWE-discovery algorithm mMERGE (Gries, 2022; Ben Youssef, 2024), resulting in a corpus that includes several thousand MWEs (MWE-BNC). Second, using the Kullback-Leibler divergence (D_{KL}) as a keyness measure (Gries, 2021; 2024), we performed two analyses of the top 1,000 keywords of BNC_{acad} vs. BNC_{rest} and MWE-BNC_{acad} vs. MWE-BNC_{rest}. Qualitatively, the comparison shows that including MWEs provides more nuanced insights, particularly in disambiguating the multiple senses keywords may have and, thus, refining domain-specific keyword categorization. Statistically, the addition of MWEs also results in significant shifts in keyword rankings. Keywords embedded within fewer MWEs tend to increase in prominence, decreasing in rank, thus becoming more distinctive of the texts they appear in and, consequently, becoming more informative. Conversely, keywords found across a broader set of MWEs often show a decline in rank, indicating less distinctive usage.

List of references

Almaghlouth, S. (2022). Environmental sustainability in the online media discourses of Saudi Arabia: A corpus-based study of keyness, intertextuality, and interdiscursivity (X. Li, Ed.). PLOS ONE, 17 (11), e0277253.

Baker, P., & McEnery, T. (2005). A corpus-based approach to discourses of refugees and asylum seekers in UN and newspaper texts. Journal of Language and Politics, 4 (2), 197–226.

Ben Youssef, C. (2024). mMERGE: A corpus-driven Multiword Expressions discovery algorithm (Doctoral dissertation). University of California, Santa Barbara. ProQuest Dissertations & Theses Global.

Egbert, J., & Biber, D. (2023). Key feature analysis: A simple, yet powerful method for comparing text varieties. Corpora, 18 (1), 121–133.

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

Gries, S. T. (2022). Multi-word units (and tokenization more generally): A multidimensional and largely information-theoretic approach. Lexis, 1 (19).

Gries, S. T. (2024). Frequency, dispersion, association, and keyness: Revising and tupleizing corpus-linguistic measures. John Benjamins Publishing Company.

Leech, G., & Fallon, R. (1992). Computer corpora – what do they tell us about culture? ICAME Journal, 16, 29–50.

Moreno-Ortiz, A., & García-Gámez, M. (2023). Strategies for the analysis of large social media corpora: Sampling and keyword extraction methods. Corpus Pragmatics, 7 (3), 241–265.

Scott, M., & Tribble, C. (2006). Textual patterns: Key words and corpus analysis in language education. John Benjamins Publishing Company.

Who wants to be a billionaire? The semantic prosody of megacapitalist class names

Diana ben-Aaron

Independent scholar (formerly Reading, QMUL)

This paper examines the discursive construction of 'billionaire' in 20th and 21st century US English and using similarity and difference keyness to contrast the attributes and activities associated with 'billionaire' and the earlier top wealth category of 'millionaire'. Like other corpus-informed critical discourse studies, it uses collocates as a path into a broader exploration of word meaning. Data were collected from COCA (1990-2019), TenTen 2015, Google Ngrams and Twitter in 2020, forming a time capsule ahead of the present political cycle.

The concordances show a dominant discourse of both billionaires and millionaires as an exceptional but natural kind of person with attributes of involvement in business and investing, formation through effort and luck (inheritance is less often mentioned), and political interest in avoiding taxation. The frequency of 'millionaire' is fairly steady, while 'billionaire' increases during the period and gains direct association with politics with the ascent of billionaire politicians such as Mitt Romney and Michael Bloomberg.

The dominant discourse competes with a growing resistant discourse, particularly for 'billionaire', marked by words like 'greedy' and 'corrupt'. The resistant discourse is more evident in the web corpus, and especially in the Twitter data, where individuals such as Senator Bernie Sanders and author Anand Giridharadas argue that billionaires have illegitimately extracted wealth from their co-citizens and should be taxed to return it.

Unlike status markers that are proxies for socioeconomic class, 'billionaire' and 'millionaire' are based directly on capital. The demonstrated culturally specific associations of these words challenge the ideology of money as a 'neutral veil' (Mooney & Sifaki, eds., 2017) for social measurement. The data could serve as a useful baseline for diachronic extension studies and comparison with other wealth categories such as 'the one percent' and 'oligarch', as well as a case study for working with less frequent category terms.

List of references

Baker, Paul. 2006. Public discourses of gay men. Taylor & Francis.
Baker, Paul. 2011. Times may change, but we will always have money: Diachronic variation in recent British English. Journal of English Linguistics 39(1): 65-88.
Baker, Paul, et al. 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 & Society 19(3): 273-306.
Brezina, Vaclav. 2018. Statistical choices in corpus-based discourse analysis. In Charlotte Taylor and Anna Marchi, eds., Corpus Approaches to Discourse. Routledge. 259-280.
Duguid, Alison and Alan Partington. 2018. Statistical choices in corpus-based discourse analysis. In Charlotte Taylor and Anna Marchi, eds., Corpus Approaches to Discourse. Routledge. 38-59.
Gabrielatos, Costas. 2018. Keyness analysis: nature, metrics and techniques. In Charlotte Taylor and Anna Marchi, eds., Corpus Approaches to Discourse. Routledge. 225-258.
Gabrielatos, Costas, and Paul Baker. (2008). Fleeing, sneaking, flooding: A corpus analysis of discursive constructions of refugees and asylum seekers in the UK press, 1996-2005. Journal of English linguistics, 36(1): 5-38.
Jaworski, Adam, and Crispin Thurlow. 2017. Mediatizing the "super-rich": normalizing privilege. Social Semiotics.27 (3): 276-287.
Jaworski, Adam, and Crispin Thurlow. 2009. Taking an elitist stance: ideology and the discursive production of social distinction. In Jaffe, Alexandra, ed. Perspectives on stance. Oxford University Press. 195-226.
Mautner, Gerlinde. 2007. Mining large corpora for social information: The case of elderly. Language in Society 36(1): 51-72.
Mooney, Annabel, and Sifaki, Evi. 2017. Snudging Cheapskates and Magnificent Profusion: The Conceptual Baggage of 'Mean 'and 'Generous'. In Annabel Mooney and Evi Sifaki, eds., The Language of Money and Debt: A Multidisciplinary Approach. Palgrave Macmillan. 105-135.
Taylor, Charlotte. 2013. Searching for similarity using corpus-assisted discourse studies. Corpora, 8(1): 81-113.

Thurlow, Crispin, and Adam Jaworski. (2017). Introducing elite discourse: The rhetorics of status, privilege, and power. Social Semiotics, 27(3): 243-254.

'I felt helpless watching the person I love become a shell of himself': A corpus-assisted discourse analysis of evaluative representations in cancer patients' caregivers' narratives

Miguel-Ángel Benitez-Castro, **Jennifer Moreno**

University of Zaragoza

Corpus methods have proven to be a powerful tool for exploring patients' first-hand experiences, which is essential to make them feel engaged in their therapeutic process and provide them with patient-centred healthcare (Semino, 2017). In the case of cancer, although patients are usually accompanied by a caregiver, who, having to provide them with both physical and emotional support, adapts their social activity and work schedule so as to fulfil the patients' needs, their role in the patient's recovery process is often undervalued, making them cope with the impact of their loved one's illness in secret (Soto et al., 2003).

Bearing the above in mind, we consider it crucial to explore cancer caregivers' construction and verbalisation of feelings and opinions in their narratives in order to understand how affected they are by their beloved one's illness so as to design strategies that foster their engagement in the patients' therapeutic process.

The present study offers a corpus-assisted critical discourse analysis of evaluative language in a corpus of narratives written by cancer patients' caregivers combining a quantitative-qualitative approach and using both Sketch Engine (Kilgarriff et al., 2014) and UAM Corpus Tool (O'Donnell, 2016) to conduct an analysis based on a psychologically-driven revised version of SFL's Appraisal Theory (Martin & White, 2005; Bednarek, 2008), with special emphasis on the affect subsystem (Fuoli, 2018).

Our findings suggest that, when narrating their beloved ones' cancer journey, English-speaking cancer caregivers tend to use medical terminology and focus on the illness itself, depriving it from their subjective and personal view. However, despite being reluctant to publicly express their feelings, caregivers seem to be willing to openly express their opinion on the various phases the patient goes through.

List of references

Bednarek, M. (2008). Emotion talk across corpora. Palgrave Macmillan

Fuoli, M. (2018). A stepwise method for annotating Appraisal. Functions of Language, 25(2), 229-258.

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

Martin, J.R., & White, P. (2005). The language of evaluation: Appraisal in English. Palgrave Macmillan.

O'Donnell, M. (2016). UAM CorpusTool. http://www.corpustool.com/

Semino, E. (2017). Corpus linguistics and metaphor. In Dancygier, B.(ed.) The Cambridge Handbook of Cognitive Linguistics (pp. 463-476). Cambridge University Press.

Soto, J., Planes, M. & Gras, M.E. (2003). Las emociones como variables relacionadas con el cambio de hábitos de salud en familiares y amigos próximos de enfermos de cáncer. Psicooncología, 1, 75-82.

Shaping LLM ideology: A multi-dimensional approach

Tony Berber Sardinha[1], Anderson Avila[2], Maria Claudia Nunes Delfino[3], Hazem Amamou[2], Rogerio Yamada[1], Ru-bing Chen[4]

[1]Pontifical Catholic University of Sao Paulo; [2]Institut National de Recherche Scientifique (INRS) / Institut National de Recherche Scientifique (INRS), Université du Québec en Outaouais (UQO); [3]São Paulo Technical College at Praia Grande / Pontifical Catholic University of São Paulo; [4]The Hong Kong Politechnic

Previous research claims large language models (LLMs) are not ideologically neutral techno-communicative systems (Bender et al., 2021). Although various methods have been used to explore LLM ideological biases, large-scale corpus-based discourse analysis is rarely applied. We conducted a Lexical Multi-Dimensional Analysis (LMDA; Berber Sardinha & Fitzsimmons-Doolan, 2025) of a corpus of all US presidential debates from 1960 through 2020, comprising 3,428 texts (ie. the candidates' responses), 596,000 words. This enabled us to model US-based political ideologies, likely present in the LLM (ie. ChatGPT) training data as they are widely disseminated in online sources. The analysis identified six discourse-based dimensions, each reflecting ideological stances. These dimensions included, e.g. 'Discourse of Equitable Prosperity through Social Responsibility vs. Discourse of Traditionalist Conservatism: Moral, Religious, and Constitutional Vigilance' (Dimension 1); and 'Discourse of Guardianship of Freedom: America's Global Mission vs. Discourse of Sovereign Libertarian Conservatism' (Dimension 2). ANOVAs comparing candidates' political party affiliations revealed minimal variation, unlike the election cycle effect, whose R-squared values reached up to 20 percent. To investigate the embedded ideology of ChatGPT and determine whether it could be influenced, we employed a Retrieval-Augmented Generation framework based on two prompt conditions. In the dimension-informed prompt, which introduced explicit ideological context into the LLM's reasoning, the input included a question related to the topics discussed by the candidates, a description of the dimension and its label, and a series of candidate responses illustrating the discourse (texts with the highest dimensional scores). And in the dimension-free prompt, the input was identical to the previous one, except it excluded the dimensional information. This allowed us to examine the LLM's default reasoning without the influence of explicit ideological framing. The responses generated under each condition were compared using LMDA to identify discursive ideological patterns, whose results will be detailed in the paper presentation.

List of references

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? presented at FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency,

Berber Sardinha, T., & Fitzsimmons-Doolan, S. (2025). Lexical Multidimensional Analysis: Identifying Discourses and Ideologies. Cambridge: Cambridge University Press.

A large-scale key feature analysis of academic writing in the humanities

Tony Berber Sardinha[1], Deise Dutra[2], Maria Claudia Nunes Delfino[3], Ana Bocorny[4], Marilisa Shimazumi[5], Carlos Kauffmann[1], Luciana Dias Macedo[2], Ana Clara Taborda[2]

[1]Pontifical Catholic University of Sao Paulo; [2]Federal University of Minas Gerais; [3]Sao Paulo Technological College; [4]Federal University of Rio Grande do Sul; [5]Cultura Inglesa College

This paper presents the first large-scale Key Feature Analysis (KFA; Egbert & Biber, 2023) comparing individual KFAs to identify recurring KF patterns. KFA is a method used to identify the most and least prominent grammatical features across different texts, following the same logic as its predecessor, Keyword Analysis (Scott, 1997). KFA compares the mean and standard deviation of each grammatical feature in a target corpus to those in a reference corpus. The effect size of the difference is calculated using Cohen's d, and features that exceed a specified cut-off point are considered key features. The primary aim of the analysis was to determine essential grammatical features that could inform the development of language teaching materials and guide course planning. The corpus consisted of 4,028 research articles (ca. 6 million words) in English. Texts were divided into sections (abstract, introduction, methods, results/discussion, conclusion), tagged using the Biber Tagger, and post-processed with the Biber Tag Count program. A three-level analysis was conducted: field, section, and field specific section. Reference corpora were dynamically built using texts excluded from each level. Cluster analysis was employed to identify KFA patterns. Findings: at the field level, a large cluster of 10 fields, and three specialized clusters were found. At the level of sections, three groups were found: (1) introductions, results/discussions, and conclusions; (2) abstracts; and (3) methods. And at the field-section level, cluster analysis of the clustering solutions revealed six patterns, the most common of which was for three clusters comprising (1) abstracts and methods, (2) introductions and results/discussion, and (3) conclusions. The clusters will be detailed and discussed in the paper presentation. The implications include enabling teachers to design lessons and materials by (a) targeting shared KFA patterns across fields and (b) focusing on features specific to each field or section.

List of references

Egbert, J., & Biber, D. (2023). Key feature analysis: a simple, yet powerful method for comparing text varieties. Corpora, 18(1), 121-133.

Scott, M. (1997). PC Analysis of key words - and key key words. System, 25(2), 233-245.

Longitudinal development of syntactic complexity in Chinese ESL MA students

Liwen Bing

University of Birmingham

In L2 writing development, syntactic complexity is a vital construct, as its development is essential to a second language learner's overall progress in mastering the target language (Ortega, 2003). Recent research views syntactic complexity as a multidimensional construct (Bulté & Housen, 2014; Norris & Ortega, 2009). However, the interpretation of measures and the dimensions they represent remain controversial. In addition, current research on syntactic complexity in L2 writing development largely relies on language assessment data (e.g., TOEFL) or focuses on EFL contexts, with limited studies using disciplinary corpora in ESL settings.

To address these gaps, this study examined the longitudinal development of syntactic complexity measures in disciplinary assignments written by Chinese ESL MA students over one academic year at a UK university. The dataset consisted of 124 academic assignments written by 62 participants, collected once per semester per learner, forming a self-built corpus. Exploratory factor analysis was conducted on 14 syntactic complexity measures calculated using the L2 Syntactic Complexity Analyzer (Lu, 2010) to identify the constructs underlying these measures. The results showed that 13 of the 14 measures grouped into two factors.

The first factor represents clausal sophistication and comprised six measures including T-unit complexity ratio (C/T) and dependent clauses per T-unit (DC/T). In contract, the second factor represents phrasal sophistication, which included seven measures, such as mean length of clause (MLC) and complex nominals per clause (CN/C). These two factors explained more than 80% of the total variability of the 13 syntactic complexity measures.

Following this, mixed-effects regression models predicted the factor scores over time, controlling for individual variation. Results revealed significant increases in clausal and phrasal sophistication. Further analysis of individual measures indicated growth in dependent clauses and complex nouns. One possible interpretation of the result is based on participants' developmental stage in academic writing.

List of references

Bulté, B., & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. Journal of Second Language Writing, 26, 42–65.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. International Journal of Corpus Linguistics, 15(4), 474-496.

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. Applied Linguistics, 30(4), 555–578.

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. Applied linguistics, 24(4), 492-518.

A corpus-assisted analysis of parenthood representations in the MIAMUL corpus: Picturebooks about migration

Maria Bîrlea

University of Salamanca

Migration has received a lot of social attention lately and this has been mirrored in the publication of many picturebooks dealing with this topic (Hope, 2008). In these picturebooks, parents frequently accompany and help their children, offering consolation or describing what their new life would be like. As has been the case in earlier studies on gender in children's literature, there is a tendency to present traditional gender roles within the familial depictions with fathers being more 'invisible' (e.g., Sunderland, 2011 and Anderson & Hamilton, 2005). This work aims to advance the corpus linguistics study of how parenthood is portrayed in literature (e.g., Geybels, 2024) by addressing gender and migrant intersectional representations of parenthood in the MIAMUL Corpus.

To investigate these representations, a critical corpus-assisted analysis (Baker & McGlashan, 2020; Partington, 2008) of 40 picturebooks that include a sample of multimodal migrants' narratives was carried out. In this corpus, parenthood has been analysed through the examination of the verbal and visual collocations, as well as the verbal exchanges between parents and children. Halliday's (1978) and Kress and Van Leeuwen's (2006) integrated approaches to multimodal texts serve as the theoretical framework for this analysis. The study aimed to explore patterns in how parents are construed as participants throughout the corpus and the processes they are involved in.

Findings indicate a preference for mothers as caregivers and Sayers, notwithstanding the corpus's even distribution of [mother] and [father] presences and the small number of picturebooks portraying children without parents (15%). Results also show that moms are more likely to be portrayed as their children's supporters since kids turn to them for support and affection. Meanwhile, the primary function of fathers is to sanction their children's behaviour. These findings appear to illustrate a stereotyped portrayal of parents' gender roles in migration-themed picturebooks.

List of references

Baker, P. & McGlashan, M. (2020). 'Critical Discourse Analysis'. In Adolphs, S. & Knight, D. (Eds.), The Routledge Handbook of English Language and the Digital Humanities (pp. 220-241). Routledge.

Geybels, L. (2024). "Weird, but lovely" A digital exploration of age in David Almond's oeuvre. In V. Joosen et al. (Eds.), Age in David Almond's Oeuvre: A Multi-Method Approach to Studying Age and the Life Course in Children's Literature (pp. 59-91). Routledge. doi: https://doi.org/10.4324/9781003369608-4

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2004). Introduction to functional grammar (3rd ed.). Arnold.

Hope, Julia (2008). "'One Day We Had to Run': the Development of the Refugee Identity in Children's Literature and its Function in Education", Children's Literature in Education, 39(4), 295-304. doi: 10.1007/s10583-008-9072-x.

Kress, G., & Van Leeuwen, T. (2006). Reading Images: The Grammar of Visual Design (2nd ed.). Routledge.

Kress, G., & Van Leeuwen, T. (2021). Reading Images: The Grammar of Visual Design (3rd ed.). Routledge.

Partington, A. S. (2008). The armchair and the machine: Corpus-Assisted Discourse Studies. Essay. In Carol Taylor Torsello, Katherine Ackerley & Erik Castello (Eds.), Corpora for university language teachers (pp. 189–213). Bern, Peter Lang.

Painter, C., Martin, J., & Unsworth, L. (2012). Reading Visual Narratives: Image Analysis of Children's Picture Books. Equinox.

Sunderland, J. (2011). Language, Gender and Children's Fiction. Bloomsbury.

Self-promotional strategies in doctoral thesis conclusions: A cross-disciplinary corpus-based analysis

Emmanuel Mensah Bonsu

The Hong Kong Polytechnic University

Doctoral theses are important genres of academic writing that mark the transition of postgraduate students to become experts in their respective fields (Bunton, 2002). However, little is known about how doctoral students construct their scholarly identity and self-promote their research in the conclusion chapter—a macro-genre that offers a final opportunity to reinforce the thesis's importance (Bunton, 2005; Deng, 2012; Trafford et al., 2014). To address this gap, this study examines self-mention as a promotional strategy in doctoral thesis conclusions across three disciplines: Applied Linguistics, Psychology, and Physics. Using a two-stage sampling method, 150 doctoral theses were selected from the eight UGC-funded Hong Kong Universities (2015–2024). Drawing on Dontcheva-Navratilova's (2023) self-mention framework, the study utilised LancsBox X (5.0.3) to search for the frequencies of grammatical forms of self-mention, UAM Corpus Tool (6.0) to annotate and calculate the frequencies of their rhetorical roles, and JASP (0.19.3) to conduct one-way between subject ANOVA analyses on the frequencies of the grammatical forms and rhetorical roles. Analyses revealed significant disciplinary differences in the grammatical forms and rhetorical roles of self-mention. Grammatically, grammatical subjects in the nominative case differed across disciplines. Rhetorically, significant disciplinary differences were found in in the use of researcher roles. Based on these findings, implications are derived for English-for-academic-purposes pedagogy, doctoral supervision, and further research on promotion.

List of references

Bunton, D. (2002). Generic moves in Ph.D. thesis introductions. In J. Flowerdew (Ed.), Academic discourse (pp. 57-75). Pearson Education.

Bunton, D. (2005). The structure of PhD conclusion chapters. Journal of English for Academic Purposes, 4(3), 207-224.

Deng, L. (2012). Academic identity construction in writing the discussion & conclusion section of L2 theses: Case studies of Chinese social science doctoral students. Chinese Journal of Applied Linguistics, 35(3), 301-323.

Dontcheva-Navratilova, O. (2023). Self-mention in L2 (Czech) learner academic discourse: Realisations, functions and distribution across master's theses. Journal of English for Academic Purposes, 64, 101272.

Hyland, K. (2004). Disciplinary interactions: Metadiscourse in L2 postgraduate writing. Journal of Second Language Writing, 13(2), 133-151.

Kwan, B. S. (2006). The schematic structure of literature reviews in doctoral theses of applied linguistics. English for Specific Purposes, 25(1), 30-55.

Starfield, S., Paltridge, B., McMurtrie, R., Holbrook, A., Bourke, S., Fairbairn, H., ... & Lovat, T. (2015). Understanding the language of evaluation in examiners' reports on doctoral theses. Linguistics and Education, 31, 130-144.

Trafford, V., Leshem, S., & Bitzer, E. (2014). Conclusion chapters in doctoral theses: some international findings. Higher Education Review, 46(3).

Linking external events to discourse shifts in Slovenian parliamentary debates

David Bordon

University of Ljubljana

The proposed work explores the relationship between significant political events and thematic shifts in parliamentary discourse by combining computational methods and a corpus-based approach.

Using a subset of the linguistically annotated Slovene parliamentary corpus siParl 4.0 (Pančur et al., 2024), covering plenary debates from 1992 to 2022, the aim of the work is to study discourse shifts around major scandals, high profile corruption cases, and other events of great relevance for Slovenia based on media coverage and historical consensus.

The focus is on showing how certain political themes persist long after a momentous event, while others briefly arise and quickly recede. The implications of the work extend to studying how specific crises can be used to mobilize polarizing discourses, by referencing “ghosts of the past”, to assign blame, revise historical narratives, and shape public opinion.

The work employs dynamic topic modeling, specifically the BERTopic algorithm based on sentence transformers, allowing for nuanced semantic clustering (Grootendorst, 2022), as a distant reading technique to define the continuity (or lack thereof) of political discourses. The validity of the topic model results will be tested by cross-referencing the findings with a diachronic corpus frequency analysis. The most relevant terms in each topic will be identified and tracked diachronically, seeking correlation with external events, referencing official records and media coverage to validate the interpretative accuracy of the topic clusters.

The proposed work offers a methodology that can be adapted to other national context (using the ParlaMint family of corpora (Erjavec et al., 2024)), while exploring how language manifests the interplay between scandals and shifts in discourse in times of political turmoil, giving insights into rhetorical strategies that might be used to influence long-term public opinion formation.

List of references

Baker, P., Gabrielatos, C., & McEnery, T. (2013). Discourse analysis and media attitudes: The representation of Islam in the British press. Cambridge University Press. https://doi.org/10.1017/CBO9780511920103.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In Proceedings of the 23rd international conference on Machine learning (pp. 113-120). https://doi.org/10.1145/1143844.1143859.

Clarke, I., Brookes, G., & McEnery, T. (2022). Keywords through time. International Journal of Corpus Linguistics, 27(4), 399-427. https://doi.org/10.1075/ijcl.22011.cla.

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure, arXiv preprint arXiv:2203.05794.

Laver, M., Benoit, K., & Garry, J. (2003). Extracting policy positions from political texts using words as data. American Political Science Review, 97(2), 311–331. http://www.jstor.org/stable/3118211?origin=JSTOR-pdf.

Meden, K., Erjavec, T. & Pančur, A. Slovenian parliamentary corpus siParl. Lang Resources & Evaluation (2024). https://doi.org/10.1007/s10579-024-09746-8.

Moretti, F. (2013). Distant Reading. Verso.

Pančur, Andrej; et al., (2024), Slovenian parliamentary corpus (1990-2022) siParl 4.0, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1936.

Rayson, P., & Garside, R. (2000). Comparing corpora using frequency profiling. In Proceedings of the workshop on Comparing Corpora (pp. 1-6). https://doi.org/10.3115/1117729.1117730.

Tomaž Erjavec et al. (2024) Multilingual comparable corpora of parliamentary debates ParlaMint 4.1. http://hdl.handle.net/11356/1912.

van Dijk, T. A. (2006). Discourse and manipulation. Discourse & Society, 17(2), 359-383. https://doi.org/10.1177/0957926506060250.

Corpus linguistics meets legal tech: A hybrid text mining approach to combat money laundering

Steffen Bothe, **Stephanie Evert**

Friedrich-Alexander-Universität Erlangen-Nürnberg

We present an ongoing research project that aims to automate legal procedures involving the German Commercial Register. Our sample comprises around 12 million documents associated with ca. 500,000 register entries for companies. This data set is analysed by a highly interdisciplinary team combining expertise from jurisprudence, corpus linguistics, natural language processing, artificial intelligence, knowledge representation, and automatic reasoning.

One of our use cases in the project is computer-assisted analysis of the shareholder structure of companies in order to combat money laundering. Due to the slow adoption of legal tech applications in Germany, this process currently still requires highly-paid notaries to sift through a database of unstructured, scanned documents in order to work out which natural persons are the ultimate economic beneficiaries of a company. As a further complication, multiple shells of proxy companies can separate a company from its true beneficiaries.

In order to automate this tedious process, we have created a corpus from the relevant documents, converting them into machine-readable text via state-of-the-art OCR tools and adding linguistic annotation such as POS tags, lemmata, and named entities. We then identify passages that describe the shareholder structures with a hybrid approach: (i) based on similarity search with sentence embeddings (Reimers & Gurevych 2019), and (ii) based on bespoke CQP corpus queries (Evert & Hardie 2011) that are developed interactively following the approach of Dykes et al. (2021). In our presentation, we compare the two approaches in terms of efficiency, accuracy, and complementarity, showing the advantages of a hybrid combination.

Once a relevant passage has been identified, named entities are extracted and the corresponding shareholder-company relationships are inserted into a triple store. After processing all available documents, the shareholder tree of any given company can be visualised and its ultimate beneficiaries can be identified by logical inference in the triple store.

List of references

Dykes, N., Evert, S., Göttlinger, M., Heinrich, P., & Schröder, L. (2021). Argument parsing via corpus queries. it – Information Technology, 63(1):31–44.

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.

Reimers, N. & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3982–3992, Hong Kong, China.

Lexical complexity in Spanish-language European Union regulations

Earl Kjar Brown, James Robinson

Brigham Young University

Legal language is difficult for laypeople to understand because of long sentences, impersonal constructions, low-frequency and arcane words and phrases, and technical terms specific to law (Tiersma 1999; Bednarek 2014). This reality creates tension, as laypeople are expected to abide by laws and regulations that they do not understand well. In response, the Plain Language Movement has the goal of making legal language easier to understand. Recently, academic studies have started to quantify the amount of linguistic complexity in legal language (Boyd & Walbaum Robinson 2015; Marasigan & Ballesteros-Lintao 2020). For example, Hashimoto, Brown and Marshall (under review) measured the level of lexical complexity in the US Code and compared it to three subcorpora of the Corpus of Contemporary American English (COCA): academic, newspaper, and television and movie subtitles.

The present study is a conceptual replication of Hashimoto, Brown and Marshall (under review), albeit an analysis of a different legal context and a different language. We downloaded all regulations of the European Union (EU) in Spanish (n = 14,617) and analyzed them for lexical complexity. Following Bulté and Housen (2012), we operationalized lexical complexity as: lexical density, lexical diversity, and lexical sophistication. Specifically, we measured lexical density as the rate of content words and nouns, lexical diversity with three modern measures (MATTR, HDD, MTLD-wrap), and lexical sophistication as the average frequency and dispersion of content words in each text. Our comparison corpora are academic articles (n = 3,139), newspaper articles (n = 6,555), and television and movie subtitles (n = 7,177). The analysis returned both expected and unexpected results. For example, as expected, regulations texts have lower-frequency content words on average, but unexpectedly, regulations texts have the least lexically diverse vocabulary among the four Spanish registers we compared. Implications for the simplification of legal language in EU Spanish-language regulations are offered.

List of references

Bednarek, Grażyna. 2014. On Lexical and Syntactic Qualities of the English Language of Law. Heteroglossia 4. 63–75.

Boyd, Michael S. & Walbaum Robinson, Isabel A. 2015. Text Commenting in Mediatised Legal Discourse: Evaluating Reader Understanding of (International) Criminal Law. International Journal of Law, Language & Discourse 5(1). 1–37.

Hashimoto, Brett, Brown, Earl Kjar and Marshall, Catherine. Under review. Lexical complexity in US statutes.

Marasigan, Michelle Anne A. & Ballesteros-Lintao, Rachelle. 2020. Presentation and Comprehensibility of Public Policies in Online News Articles. International Journal of Law, Language & Discourse 8(2). 35–56.

Tiersma, Peter M. 1999. Legal Language. Chicago: University of Chicago Press.

Beyond the brochure: Evaluating student mental health support through online university reviews

Savannah T. Brown, **Jack A. Hardy**

Oxford College of Emory University

In an era of increasing attention to student mental health, understanding how college environments support—or fail to support—student well-being is crucial. Online university reviews, a growing genre of feedback, offer unique insights into the lived experiences and perceptions of students, particularly regarding institutional support systems. These reviews not only guide prospective students in their college decisions but also provide a valuable resource for universities aiming to assess and improve student outcomes (Han, 2014). This study leverages the American Corpus of University Reviews (ACUR), a 5.8-million-word dataset derived from Niche.com, to analyze how students discuss mental health alongside broader evaluations of institutional quality. Drawing on corpus-based discourse analysis, we identify linguistic patterns related to protective and risk factors (Campbell et al., 2022) and examine how these patterns vary across institutions.

To triangulate these findings, we integrate a quantitative evaluation of institutional mental health policies using a rubric adapted from Heyman (2018). This rubric evaluates comprehensiveness, fairness, and student-centeredness in mental health policies, assigning each institution a letter grade and a numerical score. Universities are grouped into “Satisfactory” and “Unsatisfactory” policy categories, and linguistic features from student reviews, including evaluative adjectives (Edo-Marzá, 2013) and collocations, are analyzed in relation to these groupings. By correlating policy ratings with patterns in student discourse, we aim to uncover how institutional support quality influences perceptions of well-being and community.

Future directions include expanding the ACUR to incorporate temporal analyses of mental health discourse, tracking shifts in response to institutional changes or external pressures such as the COVID-19 pandemic. Comparative studies with other evaluative registers, such as online reviews of healthcare providers (Baker et al., 2019), will further refine methodologies for linking corpus-based insights to institutional practices. This research ultimately seeks to provide actionable recommendations for universities striving to enhance their mental health support systems.

List of references

Baker, P., Brookes, G., & Evans, C. (2019). The language of patient feedback: A corpus linguistic study of online health communication. Routledge. https://doi.org/10.4324/9780429259265

Campbell, F., Blank, L., Cantrell, A., Baxter, S., Blackmore, C., Dixon, J., & Goyder, E. (2022). Factors that influence mental health of university and college students in the UK: A systematic review. BMC Public Health, 22(1), 1–22. https://doi.org/10.1186/s12889-021-12380-9

Edo-Marzá, N. (2013). The formation of the image of top-ranked hotels through real online customer reviews: A corpus-based study of evaluative adjectives as image-formers/providers. Journal of Pragmatics, 45(1), 67–88. https://doi.org/10.1016/j.pragma.2012.11.007

Han, P. (2014). A literature review on college choice and marketing strategies for recruitment. Family and Consumer Sciences Research Journal, 43(2), 120-130.

Heyman, M. (2018). The Ruderman white paper on mental health in the Ivy League. Ruderman Family Foundation.

Rethinking relative clauses in the light of the Spoken BNC2014: From prototypical to frequent

Graham Burton[1], Christian Jones[2]

[1]Free University of Bozen-Bolzano; [2]University of Liverpool

Prototype theory posits that people 'define a concept by reference to typical instances' (Richards & Schmidt, 2010, p. 471), but these instances are not necessarily the most frequent. In this paper, we apply the concept of prototypicality to descriptions of grammar. Drawing on insights from a recently-published book on conversational English (anonymised reference) based on data from a conversational corpus (specifically, the Spoken BNC2014), we explore the question of how a corpus-based, frequency-informed approach to grammar analysis can lead to a rethink of some standard, prototypical explanations of English grammar.

We focus here on relative clauses, descriptions of which are typically structured around the prototypical contrast between restrictive and non-restrictive clauses and tend to provide concrete, prototypical examples of them to unequivocally illustrate their function. Firstly, we show how data from the Spoken BNC2014 shows that this distinction is in fact hard to maintain when conversational data is considered. We also argue that the corpus data shows that in any case it is often not appropriate to explain the use of the relative clause as 'restricting' the scope of the antecedent – as in conventional descriptions based on prototypical uses – since the role of the latter, at least in conversational English, appears often to be no more than a syntactic place holder.

We argue that while it is not unreasonable for grammatical descriptions to focus on prototypical uses, they should not be dominated by explanations and examples that are not necessarily borne out in empirical data. Taking a bottom-up, frequency-based approach can reveal uses that while perhaps not prototypical are nonetheless key for gaining a true understanding of conversational discourse.

List of references

[anonymised reference]
Richards, J. C., & Schmidt, R. W. (2010). Longman dictionary of language teaching and applied linguistics (4th ed.). Longman.

Adapting and applying USAS tagging to medieval Scots

Megan Elizabeth Bushnell

University of Oxford

Medieval Scottish texts present certain challenges for corpus linguistic analysis. Medieval Scots encompasses a variety of lexis and orthography that can complicate tagging and annotation. There are no existing tools for the automated tagging of medieval Scottish texts, and even normalisation tools like VARD require adaptation for use with Scots. Moreover, since the existing corpus of medieval Scottish texts is not large, it can be difficult to develop such tools. Nevertheless, these works are productive for corpus linguistic study, especially for understanding the evolution of Scots and its relationship to English, which is fraught with political significance. This paper uses a corpus linguistic project analysing Gavin Douglas's Eneados – an Older Scots translation of the Aeneid written in 1513 – as a case study for adapting available tagging tools for use on medieval texts written in Scots.

This paper reflects on the process of annotating and tagging the text using tools that were designed for modern English – namely, the USAS tagger. It describes the semi-automated process of normalising the text and 'translating' it into modern English, the automated process of running it through the USAS tagger, and then the manual process of reviewing and editing the tagging. It discusses the practical and conceptual challenges of this work, its successes, and its failures. This paper concludes with a set of recommendations for how to adapt tools developed for modern languages for historical or minority languages as a tentative step towards best practices. It also looks to the future for how this work could be further developed to benefit future work done on Older Scots, and how it could be integrated with other work in this field, such as adaptations of USAS for other older varieties of English.

Mapping urban identities: A corpus approach to discursive place-making in Brooklyn, New York

Beatrix Busse, **Nina Dumrukcic**

University of Cologne

This study, part of a project called heiUrban, investigates how residents and visitors in Brooklyn, New York engage in dynamic discursive place-making, focusing on the ways they construct and negotiate their diverse and evolving identities in an urban setting marked by rapid demographic and cultural shifts.

Discursive place-making and semiotic meaning-making are interconnected processes that shape both the experience of and interaction with urban environments. The transformation of space into place involves negotiating the diverse voices, signs, symbols and interests converging within the city (Busse 2019, 2021, 2022; Cresswell 2015; Busse & Warnke 2022).

The heiUrban corpus is compiled by collecting and organizing language data from various modalities such as text, speech, image, and video within one digital interface. This study focuses specifically on the sub-corpus of 1579 semi-structured interviews (transcriptions amount to ca. 536,923 tokens including annotations) and analysis of Wi-Fi SSID (Service Set Identifier) names collected in 2017 (46,903 individual SSIDs) and 2023 (57,646). Our aim is to show how individuals in Brooklyn encode cultural, social, and geographic meanings into both verbal narratives and digital markers in areas undergoing gentrification.

We discuss the challenges of multimodal corpus data (Mana et al. 2007; Knight and Adolphs 2020) and the role of the researcher in urban ethnography (Pink 2008; Busse and Warnke 2022). Studies on mapping urban linguistic diversity (Craig et al. 2022; Väisänen et al. 2022) have focused on detecting which languages are used in urban places, however, there is a lack of research on lexico-grammatical and sociopragmatic features used in these urban environments, and we will present these findings in our presentation.

By combining standard corpus analyses such as keyness, concordances, and collocational networks with geo-mapping and NLP techniques i.e. sentiment analysis, this study serves as a bridge between corpus, computational and the emerging field of urban linguistics.

List of references

Busse, B. (2019). Patterns of discursive urban place-making in Brooklyn, New York. In V. Wiegand & M. Mahlberg (Eds.), Corpus linguistics, context and culture (pp. 13–42). Berlin: Mouton de Gruyter.

Busse, B. (2021). Practices of discursive urban place-making in Brooklyn, New York: (Hidden) digital and embodied discourse. Text & Talk, 41(5-6), 617–641.

Busse, B. (2022). The HeiURBAN Database: A brief and unconventional position piece. In B. Busse & I. Warnke (Eds.), Handbuch Sprache im urbanen Raum / Handbook of language in urban space (pp. 394–414). Berlin & Boston: De Gruyter.

Busse, B., & Warnke, I. (2022). Urban linguistics: Ideas and anchor points. In B. Busse & I. Warnke (Eds.), Handbuch Sprache im urbanen Raum / Handbook of language in urban space (pp. 1–32). Berlin & Boston: De Gruyter.

Craig, S., Daurio, M., Kaufman, D., Lampel, J., Perlin, R., & Turin, M. (2022). Mapping urban linguistic diversity in New York City: Motives, methods, tools, and outcomes. Language Documentation and Conservation, 15, 458–490.

Cresswell, T. (2015). Place: An introduction (2nd ed.). Malden: Wiley Blackwell.

Knight, D., & Adolphs, S. (2020). Multimodal corpora. In M. Paquot & S. T. Gries (Eds.), A practical handbook of corpus linguistics (pp. 1–16). Springer, Cham.

Mana, N., Lepri, B., Chippendale, P., Cappelletti, A., Pianesi, F., Svaizer, P., & Zancanaro, M. (2007). Multimodal corpus of multi-party meetings for automatic social behavior analysis and personality traits detection. In Proceedings of the Workshop on Tagging, Mining, and Retrieval of Human-Related Activity Information at ICMI'07 (pp. 9–14). Nagoya, Japan.

Pink, S. (2008). An urban tour: The sensory sociality of ethnographic place-making. Ethnography, 9(2), 175–196.

Vaisanen, T., Jarv, O., Toivonen, T., & Hiippala, T. (2022). Mapping urban linguistic diversity with social media and population register data. Computers, Environment and Urban Systems, 97.

Reconstructing spoken languages in the past: A study of multi-word constructions in the Old Bailey Corpus and a Diachronic Corpus of Indian English (DiCIE)

Rita Calabrese

University of Salerno

The study of spoken languages at time periods predating the application of recording tools is a challenging issue in the field of historical linguistics. The availability of data close to real speech such as trial proceedings recorded in the Old Bailey and The Statesman (including recordings of legal cross examinations, <LCE>) provides the opportunity to analyse spoken data dating back to Late Modern English and evaluate the degree of stabilization over time in Indian English, which is today considered "one of the most widespread and abundantly used varieties of English [..] in virtually every corner of the globe" (Lambert 2018). Drawing on both the sociolinguistic model of contact language development and the longitudinal model of linguistic structural reanalysis, this preliminary study aims to investigate the structure of multi-word constructions (MWCs) in IE and verify whether they have undergone structural reanalysis over time. The corpora under study include <LCE> dating back to 1909 extracted from the Diachronic Corpus of Indian English (DiCIE) specifically compiled at the University of Salerno for a diachronic investigation covering a period of 100 years and parallel selected sections of the Old Bailey Corpus. The two corpora were automatically annotated using the VISL interface and the UCREL Semantic Annotation System (USAS) and then compared with similar data from ICE-IND. The analysis was carried out to test two hypotheses: 1. Data dating back to different time periods show a number of shared traits of convergence toward the sets of localised forms. 2. Data show clear signs of divergence with respect to a. the target norms of British English, b. the set of localised forms identified in past literature. The automatic procedure adopted to annotate and extract data has shown that the most discriminative features in determining the choice of given constituents in MWCs rely on semantic factors.

Multimodal meaning-making profiles in L2 writing: A corpus approach

Duygu Candarli

University of Southampton

Multimodal communication is increasingly important in digital writing, partly due to rapidly evolving technological advances. Despite these trends in multimodal communication, multimodality, especially imagery, remains an oversight in corpus linguistics (CL) research, apart from notable exceptions in corpus-assisted discourse analysis studies (e.g., Baker & Collins, 2023; Christiansen et al., 2020). In second language (L2) writing research, multimodality has been investigated mostly within a case-study design, limiting generalisability. This lack of cross-fertilisation between CL and theoretical frameworks of multimodality has hindered empirical advancement in the field. This study bridges these gaps and addresses the following research questions: What are the multimodal meaning-making profiles in L2 discipline-specific writing? What is the relationship between the multimodal meaning-making profiles and the writer-related factors, including age, gender, discipline and language proficiency? This presentation uses a balanced corpus of 100 successful multimodal discipline-specific writing produced by L2 writers of postgraduate students in four disciplinary groups in UK higher education. A 'modest' XML (extensible markup language) annotation (Hardie, 2014) was utilised to tag various multimodal features, including tables, text boxes and images (over 600), which had mostly been treated as 'noise' in previous corpora, and their descriptions. A corpus tool, LancsBox X (Brezina & Platt, 2024), was used to examine the semantic representations of multimodal features in each text. Then, a cluster analysis was employed to reveal distinct multimodal meaning-making profiles in L2 writing, and the rhetorical structure theory extended to multimodal communication (Bateman, 2008) was used as a lens to interpret the findings, bridging the fields of CL and multimodality. The findings revealed the clusters of evaluation, cause-effect and circumstance (time and location) meaning relations conveyed through multimodal features, which would otherwise have been missed. The methodological affordances and limitations of treating multimodal features as 'text' will be discussed for CL research, and the pedagogical implications of findings will be outlined.

List of references

References

Baker, P., & Collins, L. (2023). Creating and analysing a multimodal corpus of news texts with Google Cloud Vision's automatic image tagger. Applied Corpus Linguistics, 3(1), 100043. https://doi.org/10.1016/j.acorp.2023.100043

Bateman, J. A. (2008). Multimodality and genre: A foundation for the systematic analysis of multimodal documents. Palgrave Macmillan.

Brezina, V. & Platt, W. (2024). #LancsBox X [software], Lancaster University, http://lancsbox.lancs.ac.uk.

Christiansen, A., Dance, W., & Wild, A. (2020). Constructing corpora from images and text: an introduction to visual constituent analysis. In S. Rüdiger & D. Dayter (Eds.), Corpus approaches to social media (pp. 149–174). John Benjamins.

Hardie, A. (2014). Modest XML for corpora: Not a standard, but a suggestion. ICAME Journal, 38, 73–103. https://doi.org/10.2478/icame-2014-0004

Introducing the CLARUS corpus: Addressing ambiguity in digital forensics across organisations and borders

Duygu Candarli[1], James Balfour[2]

[1]University of Southampton; [2]University of Glasgow

'CLARUS – building clarity and preventing bias in digital forensic examinations' is a multidisciplinary and multi-organisational project that addresses ambiguous language use in digital forensic and investigative documents in five European countries — Finland, Czechia, Portugal, Greece and the UK. Although there is a growing use of corpora and corpus methods in forensic and legal linguistics (e.g., Wright, 2025), corpora are often limited to one national context, and corpora containing sensitive documents are rare. This study addresses these challenges by collecting an 800,000-word corpus representative of written digital forensic and investigative documents in five European countries. The aim of this presentation is three-fold: (1) We illustrate our innovative approach to corpus construction that involves manually tagging all the multimodal features, including images, tables and typographical features (italics, etc.) in documents and inter-annotator agreement on these; (2) We discuss challenges of working in partnership with practitioners, including police, forensic scientists, and academics across disciplines like criminology and psychology, as well as issues with translation, metadata collection, and our strategies to overcome them; (3) We present a case study examining semantic domains in digital forensic communication across five countries. The corpus of this Horizon Europe-funded project consists of legislation, manuals, reports, and codes of practice that are distributed among police officers and forensic scientists in five countries. Using LancsBox X (Brezina & Platt, 2024), the semantic domains of language use were examined utilising the USAS semantic tag sets, and we read randomly selected concordance lines to interpret these semantic representations. Our preliminary findings indicate cross-cultural differences in the use of certain semantic domains, such as evaluation and modality. We discuss the implications of our findings, plans for further research, engagement and impact.

List of references

References

Brezina, V. & Platt, W. (2024). #LancsBox X [software], Lancaster University, http://lancsbox.lancs.ac.uk.

Wright, D. (2025). Corpus approaches to discourse in forensic and legal contexts. Routledge.

Finetuning the descriptors of the CEFR: A contrastive analysis of flexibility

María Luisa Carrió-Pastor

Universitat Politècnica de València

The Common European Framework of Reference for Languages (CEFR, 2020) identifies a common metalanguage that encompasses the main aspects related to language teaching, learning, and assessment (North, 2021). Thus, it promotes reflection on learners' needs, sets objectives and levels, and identifies ways to follow up and check their progress. These aims are encapsulated in the so-called can-do descriptors, which may be perceived as too global to provide a linguistic description of the performance of English learners in the different competences (Hawkins & Filipović, 2012). Specifically, the need to complement pragmatic descriptors with detailed information may help learners understand what is expected from them at each CEFR level. The objective of this analysis is to identify the strategies used in B2 and C1 levels in one of the descriptors of the pragmatic competence: flexibility. The corpus used in this analysis was compiled in the FineDesc project: opinion essays and emails produced by Spanish learners of English with B2 and C1 levels. Thus, the corpus was studied and all the pragmatic strategies were identified. In the results, the strategies were classified considering the use of metadiscourse devices that infer flexibility to discourse, such as Boosters, reformulators, hedges and sentence length was also added. Some examples were provided and the devices used to infer flexibility by learners with B2 and C1 English levels were contrasted. In conclusion, some causes that may interfere with the use of flexibility devices were discussed, such as the influence of Spanish.

List of references

Council of Europe. (2020) Common European Framework of References for Languages. Companion Volume. Publications of the Council of Europe.

Hawkins, J. A. and Filipović, L. (2012). Criterial Features in L2 English. Cambridge University Press.

North, B. (2021). The CEFR Companion Volume—What's new and what might it imply for teaching/learning and for assessment? CEFR Journal. Research and Practice, 4, 5-24. https://doi.org/10.37546/JALTSIG.CEFR4-1

The role of frequency and sequentiality in shaping semantic similarity and linguistic generalization

Alvin Cheng-Hsien Chen

National Taiwan Normal University

Usage-based grammar emphasizes the role of language use in shaping grammar. A core hypothesis posits that frequency facilitates the holistic processing of linguistic units, from individual words to complex multiword combinations. Stronger sequential relationships between words (i.e., word co-occurrences) are also crucial, enhancing the processing of multiword units as cohesive chunks. These chunks form the foundation of complex grammatical structures. Crucially, the transition from concrete chunks to abstract schemas relies on the observation that recurrent chunks often exhibit shared structural and semantic patterns, enabling cognitive processes like analogy to drive the formation of abstract generalizations (cf. Bybee 2002, Diessel 2023). This study investigates this link between language use (frequency and sequentiality) and linguistic generalization. We address two key questions: (a) Do high-frequency combinations exhibit greater semantic similarity than low-frequency ones? (b) Do stronger lexical associations lead to higher semantic similarity among these combinations?

We analyzed recurrent five-word units from a 185-million-word corpus of Taiwan Mandarin. For RQ1, we employed a frequency-based approach from lexical bundle research, categorizing five-word bundles across frequency bins. For RQ2, we introduced a quantitative method to identify significant recurrent multiword units (RMUs). This method utilized transitional probabilities (TP) to measure initial and final word predictability within the multiword contexts. Sequences with consistently increasing TP values as contextual information (e.g., one to four words) expands, both forward (final word prediction) and backward (initial word prediction), were identified as cohesive chunks (see Figure 1). Semantic similarities among these units were assessed using pretrained large language models, leveraging pairwise cosine similarity as a metric.

Our findings demonstrate that high-frequency and strongly associated chunks exhibit higher semantic similarity values, supporting the hypothesis that frequency and lexical associations are crucial for the emergence of abstract linguistic structures. These results provide empirical evidence for how language use shapes linguistic abstraction.

List of references

Bybee, J. (2002). Sequentiality as the basis of constituent structure. In T. Givón & B. F. Malle (Eds.), The evolution of language out of prelanguage (pp. 109-134). Amsterdam: John Benjamins.

Diessel, H. (2023). The constructicon: Taxonomies and networks. Cambridge: Cambridge University Press.

A systematic review of academic word lists

Chen Chen

Xi'an Jiaotong-Liverpool University

Academic vocabulary lists are widely used to support students' English for Academic Purposes (EAP) vocabulary acquisition. Prominent lists, such as the Academic Word List (Coxhead, 2000), have been extensively utilised in EAP pedagogy. Despite the growing body of research on the development of academic and disciplinary vocabulary lists, there remains a need for a systematic review of these studies. This study addresses this gap by evaluating 98 studies on general academic and discipline-specific word lists, following the PRISMA guidelines for systematic reviews (Page et al., 2021). The findings reveal a significant increase in the number of publications over the past two decades, a wide range of disciplines covered by the word lists, and a predominant use of word families as the unit of counting (45% of studies). Several issues in the development of these lists are identified, including the reliance on small corpora (50% of studies used corpora with fewer than 2 million tokens), a strong predominance of written corpora (89% of studies) over spoken ones, and limited consideration of semantic meaning in word selection. Additionally, this paper discusses methodological challenges in list development and provides recommendations for future research to enhance the pedagogical effectiveness of vocabulary lists in EAP acquisition.

List of references

Coxhead, A. (2000). A New Academic Word List. TESOL Quarterly, 34(2), 213–238. https://doi.org/10.2307/3587951

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. International Journal of Surgery, 88. https://doi.org/10.1016/j.ijsu.2021.105906

Exemplification across part-genres in linguistics research articles: A local grammar based investigation

Jingyao Chen

Sichuan International Studies University

The part-genres of research articles fulfill distinct communicative functions, contributing to the overall structure and coherence of the text. Previous studies have examined differences across part-genres in terms of syntactic complexity, multiword units, and citation practices, yet relatively few investigations have focused on discursive strategies across part-genres. Exemplification, a key strategy used to clarify abstract concepts and support arguments, is particularly central to academic writing. This study adopts a local grammar approach to examine variations in exemplification across part-genres, aiming to address the following questions: (1) How does the distribution of local grammar patterns of exemplification vary across part-genres? (2) How does the strategic use of exemplification differ across part-genres?

Focusing on the structural pattern of ILM[RD]C (Introduction, Literature Review, Method, Results and Discussion, Conclusion) in Linguistics research articles, this study analyzed a corpus of 322 articles published in *Applied Linguistics* (2012–2021). Exemplification instances were retrieved using eight markers (e.g., *for example, for instance, such as, an example of*) and analyzed through local grammar terminologies. Results revealed that the use of exemplification varies significantly across part-genres, aligning with their specific communicative functions. Notably, exemplification by presenting subcategories was more frequent than exemplification through citing other studies across all part-genres. The findings point to the usefulness of local grammar in exploring part-genre distinctions and have implications for EAP writing research and pedagogy.

Fighting fraud: Corpus-assisted approaches to understanding and disrupting fraud activity on the dark-web

Emily Chiang

Aston University

In 2023, an estimated $1 trillion was stolen by scammers globally (World Economic Forum, 2024). The staggering prevalence of fraud across the world is in part attributed to the rise and evolution of online technologies, including illicit marketplaces and crime-focused discussion fora on the dark-web. Such spaces are a key affordance to fraudsters as they allow like-minded users to discuss methods and practices surrounding fraud activities while providing a level of anonymity that makes policing them very difficult. Yet, these fora are also fruitful sites for linguistic exploration regarding the behaviours and activities of groups interested in fraud.

In this paper, I report on a corpus-assisted discourse analysis (Baker, 2006) of group discussions across fraud-focused dark-web fora as a means of better understanding the nature and characteristics of these communities. Using textual data scraped from online fora representing 300-9000 usernames, I present findings from ten corpora between 40,000 and 1.5m words, focusing on key words, lexical verbs and 5-grams to examine prominent topics and discursive practice. Findings highlight high levels of user expertise and a general openness to advice-giving, pointing to the efficacy and efficiency of dark-web fora as spaces for criminal upskilling.

I discuss the implications of these findings in relation to two combative approaches to fraud; first, from a law-enforcement perspective, describing how a linguistically informed understanding of online fraud communities' interactions can assist the undercover policing of dark-web fraud fora regarding the specific task of community infiltration. Second, from a commercial perspective, demonstrating how corpus analytic methods can inform online tools designed to help commercial entities monitor dark-web spaces for fraud activity related to their products.

List of references

Baker, P. (2006). 2006. Using Corpora in Discourse Analysis. London: Continuum.

BDO (2024). Fraudtrack 2024 Report. [Online]. Available at: https://www.bdo.co.uk/en-gb/insights/advisory/forensic-services/fraudtrack#form. Accessed: May 2024.

Federal Trade Commission (2024). As Nationwide Fraud Losses Top $10 Billion in 2023, FTC Steps Up Efforts to Protect the Public. [Online]. Available at: https://www.ftc.gov/news-events/news/press-releases/2024/02/nationwide-fraud-losses-top-10-billion-2023-ftc-steps-efforts-protect-public. Accessed: May 2024.

Learner corpora and metaphors: Underlying discourses and AI-based detection

Amanda Chiarelo Boldarine

Pontifical Catholic University of Sao Paulo

In this paper, we look at the extent to which L2 speakers employ metaphors to express abstract concepts in a foreign language. The goals of the current study therefore are (1) to measure the use of metaphors in student writing in EFL; (2) to detect the discourses where the metaphors are embedded; (3) to verify the extent to which Artificial Intelligence can be used for metaphor detection. To this end, a corpus of student compositions was collected, totaling 450 texts written by young learners of English (aged 11-15), based on a range of topics. We ran a Lexical Multidimensional Analysis (LMDA; Berber Sardinha & Fitzsimmons-Doolan, 2025), which detected eight dimensions of variation across the texts. For instance, in Dimension 1 "Abstract, theoretical, and scientific knowledge versus Family dynamics, personal resilience, and emotional growth", the discourse in the positive pole emphasizes structured research and scientific analysis, while the negative one focuses on personal narratives, family relationships, and overcoming personal problems. The lexical dimensions served as a short list of word candidates for the metaphor detection task, and we looked for metaphors among abstract nouns loading in the dimensions. We employed the Metaphor Identification Procedure (MIP; Pragglejaz Group, 2007) to manually annotate the metaphors, and 245 occurrences (tokens) of candidates were annotated as metaphorical, corresponding to 37% of the candidates (types). Regarding the automatic annotation of metaphors via ChatGPT4, we tried out a range of prompts, charting the process of prompt development, such as restricting the context in which linguistic metaphor occurs. The results showed that one in every four (25%) of the metaphors identified by human analysts using the MIP protocol (Pragglejaz Group, 2007) was identified automatically through AI, which suggests that metaphor detection in corpora cannot be reliably performed by ChatGPT presently.

List of references

BERBER SARDINHA, T.; FITZSIMMONS-DOOLAN, S. (2025). Lexical Multidimensional Analysis. Cambridge: Cambridge University Press.

PRAGGLEJAZ GROUP. MIP: a method for identifying metaphorically used words in discourse. Metaphor and Symbol. 22:1, 1-39, 2007.

Commenting fast and slow: Speed of response and impoliteness in online newsreader comments

Ben Clarke, **Amelie Klamm**

University of Gothenburg

Communicative transgressions online can be more or less problematic. Developing Papacharissi's (2004) distinction between 'civility' and 'politeness' vis-à-vis 'incivility' and 'impoliteness', a fundamental distinction between two broad types can be speculated: serious communicative transgressions which put the principles of democracy at risk (e.g. targeting users according to protected characteristics, such as race) and less serious behaviours that are a nuisance (e.g. insults) but do not sacrifice democratic values.

In a recent study (Clarke & Thompson, under review), we hypothesised that uncivil responses in commenting sections were likely to be made more quickly. Using a dataset of 38 million comments from The Guardian Online, we empirically tested our hypothesis by comparing the commenting speed of all one million comments removed by The Guardian moderators for breaking their community standards (The Guardian, 2009), with the commenting speed of all 37 million comments still visible on the site. Our results showed a statistically significant relation between commenting time for blocked versus visible comments, including post-hoc checks to control for bot and commercial activity.

In this paper, we study in the same dataset textually manifest communicative transgressions of the less severe kind, doing so informed by Impoliteness Theory. We again accounted for commenting time to see if such lesser communicative transgressions also evidenced the same relationship with time as more severe uncivil types, or if they were distinct, supporting Papacharissi's (2004) distinction between im/politeness and in/civility. We developed complex corpus query strings to recall four of Culpeper's (2016) types of impoliteness trigger: 'insults', 'dismissals', 'silencers' and 'message enforcers'. These produced close to 5,000 impoliteness examples. Our results show that comments containing impoliteness are not statistically significantly quicker than comments without, with one exception: impoliteness cases having third-party targets (i.e. targeting persons outside the interaction, such as celebrities; see also Coltman-Patel et al., 2022).

List of references

Clarke, B.P. & Thompson, W.H. (under review, 2025). Fast and Furious: Temporal patterns of incivility in online comments. Submitted to New Media & Society.

Coltman-Patel, T., Dance, W., Demjén, Z., Gatherer, D., Hardaker, C., & Semino, E. (2022). 'Am I being unreasonable to vaccinate my kids against my ex's wishes?' – a corpus linguistic exploration of conflict in vaccination discussions on Mumsnet talk's AIBU forum. Discourse, Context & Media, 48,

Papacharissi, Z. (2004). Democracy online: civility, politeness, and the democratic potential of online political discussion groups. New Media & Society, 6(2) 259-283.

The Guardian. (2009). Community standards and participation guidelines. Available at: https://www.theguardian.com/community-standards

The representation of genetically modified organisms across websites known to promote pseudoscience and conspiracy theories

Isobelle Clarke[1], Kevin Gerigk[2]

[1]Lancaster University; [2]Aston University

Food security – producing enough food for the global population - is a key global challenge of this century (The Royal Society, 2009). Threats to food security include: the impacts of climate change, an increasing global population, changing consumption patterns, and the scarcity of land and water (The Royal Society, 2009). Genetic modification – altering the genes of organisms and crops to change their characteristics - is being explored as one potential solution to this challenge, such as the development of drought-tolerant crops (Waltz, 2014).

Genetic modification (GM) is a controversial topic and not all public discussions surrounding GM have been informed by independent scientific evidence (Ramakrishnan, 2016). Because GM is a method and not a product, GM crops can differ considerably, meaning that blanket statements about whether GMOs are good or bad are scientifically impossible (Ramakrishnan, 2016). Research has shown that conspiracist ideation – the tendency to endorse conspiracy theories – predicts opposition to GM food (Lewandowsky et al., 2015). Here, we analysed how websites known to promote pseudoscience and conspiracy theories talk about GMOs.

Specifically, we applied Keyword Co-occurrence Analysis (Clarke et al., 2021) to a corpus of texts from websites known to promote pseudoscience and conspiracy theories that mention GMOs. The analysis revealed 8 dimensions of keyword variation, which were interpreted for discourse, register, topic, style, and attitude towards GMOs. We found that common patterns address GMOs from the perspective of food production and consumption, with implications for health (e.g. cancer, obesity), alimentation (e.g. recipes for home-made GMO-free food), nutrition (supplement use), the use of land (moral issues), the transparency of production processes (customers being lied to), and government regulations (or lack thereof). The majority of the texts express an anti-GMO stance. We discuss these results drawing comparisons with other anti-science discourses.

List of references

The Royal Society (2009). Reaping the Benefits: Science and the sustainable intensification of global agriculture. London: The Royal Society.

Waltz, E. (2014). Beating the heat. Nature Biotechnology 32: 610-613.

Ramakrishnan, V. (2016) Foreword. In The Royal Society (eds) GM Plants: Questions and answers, p. 5. London: The Royal Society.

Lewandowsky, S., Gignac, G. E., & Oberauer, K. (2015). Correction: The Role of Conspiracist Ideation and Worldviews in Predicting Rejection of Science. PLOS ONE 10(8): e0134773

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.

Using data-driven learning to improve communicative and intercultural competence in the EFL classroom

Caroline Collet[1,2], **Stefan Diemer**[3]

[1]Saarland University; [2]University of Hamburg; [3]Trier University of Applied Sciences

The aim of this paper is twofold: (1) It introduces a new corpus, TaCoCASE, and (2) it demonstrates how it can be used in English language teaching to improve communicative and intercultural competence.

Data-driven learning (DDL) means using corpora for pedagogical purposes (Gilquin & Granger 2010) and is considered a useful method putting the learner into direct contact with authentic data (Johns 2002). While DDL is mainly used for improving communicative competence, using corpora to improve intercultural competence is to date quite rare (Fahey Palma 2022). DDL supports discursive intercultural learning (Hallet 2008) when using examples from intercultural encounters in the classroom.

The corpus used for this paper is TaCoCASE (Collet 2023), which consists of spoken computer-mediated conversations (CMC) between English-speaking students from Britain, Germany and the United States, including visual and audio data and annotated transcriptions. Via the WebCorpLSE online interface, the corpus is freely available for educational and academic purposes and is thus very well suited for application in the EFL classroom.

Following a corpus-based discourse analysis, we investigate how (new) cultural concepts are being negotiated. Using both quantitative and qualitative methodologies, we collect and present corpus examples that show which communicative strategies the participants used to explain new concepts and which common discourse structure underlies the mediation of these concepts. The examples can be used in the EFL classroom to teach intercultural (communicative) competence (Byram 1997) as per the Common European Framework of Reference (CEFR). In addition, we will present hands-on examples and lesson plans for an EFL classroom setting.

In combining DDL and the intercultural corpus examples, this paper not only provides new material that serves as a rich source for EFL teaching, it also suggests a new method and a new mindset.

List of references

Byram, M. (1997). Teaching and Assessing Intercultural Communicative Competence. Clevedon: Multilingual Matters.

Fahey Palma, T. (2022). Corpora for teaching culture and intercultural communication. In R.R. Jablonkai, & E. Csomay (Eds.), The Routledge Handbook of Corpora and English Language Teaching and Learning (pp. 116-130). New York: Routledge.

Hallet, W. (2008). Diskursfähigkeit heute. Der Diskursbegriff in Piephos Theorie der kommunikativen Kompetenz und seine zeitgemäße Weiterentwicklung für die Fremdsprachendidaktik. In M. Legutke (Ed.), Kommunikative Kompetenz als fremdsprachendidaktische Vision (pp. 76-96). Tübingen: Gunter Narr Verlag.

Gilquin, G., & Granger, S. (2010). How can data-driven learning be used in language learning? In A. O'Keeffe & M. McCarthy (Eds.), The Routledge Handbook of Corpus Linguistics (pp. 359-370). New York: Routledge.

Johns, T. (2002). Data-driven learning: the perpetual challenge. In B. Kettmann & G. Marko (Eds.), Teaching and Learning by doing corpus analysis (pp. 107-117). Amsterdam: Rodopi.

TaCoCASE. (2023). Transatlantic Component of the CASE project [Corpus]. Collet, C. Saarland University & Trier University of Applied Sciences. http://umwelt-campus.de/case/tacocase (Accessed 13 March 2024).

Updating the international corpus of English for the 21st century: Towards a standardized XML-compliant markup

Sophia Conrad[1], Stella Neumann[2], Florian Frenken[2], Gerold Schneider[1]

[1]University of Zurich; [2]RWTH Aachen University

The International Corpus of English (ICE) [2] is a well-known resource that includes spoken and written texts from multiple varieties of English (components). However, the original corpus markup has several issues, including the

presence of non-corpus material in the corpus text, misspelled and malformed tags, and tags containing metacharacters (e.g. &) [4]. Furthermore, most components were encoded using an SGML-like markup [5] ill-suited for state-of-the-art corpus research. This paper reports on an effort to update this markup to a standardized, fully XML-compliant format to preserve its original value and support continued relevance for future research. We reformat the files of nine ICE components, namely Canada, Great Britain, Hong Kong, India, Ireland, Jamaica, New Zealand, the Philippines, and Singapore. To this end, we use component-specific Python scripts which unify tag variants, replace unsuitable tags and move non-corpus text into attributes within tags while keeping most of the information that is present in the original markup. This is relevant for a detailed qualitative inspection of corpus texts while ensuring that each component is represented as well-formed XML that can be validated with a corresponding XML Schema Definition (XSD) file, which provides a specification of the document structure and markup used. Furthermore, it aligns with the requirements of corpus tools such as the IMS Open Corpus Workbench (CWB) [1] and Dependency Bank [3] geared toward quantification. The update harmonizes markup across components and thus facilitates comparative studies of varieties. By publishing scripts that restructure the components, we allow researchers to apply them to the corpora and customize the transformations as needed. As an example, an additional version is published where only a subset of the tags such as metadata (e.g. speaker IDs) and some basic formatting such as header and paragraph tags is kept along with the corpus text.

List of references

[1] Stefan Evert and Andrew Hardie. “Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium”. In: Proceedings of the Corpus Linguistics 2011 conference. Citeseer. 2011, pp. 1–21.
[2] Sidney Greenbaum and Gerald Nelson. The international corpus of English (ICE) project. 1996.
[3] Hans Martin Lehmann and Gerold Schneider. “BNC Dependency Bank 1.0”. In: Studies in Variation, Contacts and Change in English, Volume 12: Aspects of corpus linguistics: compilation, annotation, analysis. Ed. by Signe Oksefjell Ebeling, Jarle Ebeling, and Hilde Hasselg˚ard. Helsinki: Varieng, 2012.
[4] Stella Neumann. “Applying register analysis to varieties of English”. In: Anglistentag 2011 Freiburg proceedings (2012), pp. 75–94.
[5] Deanna Wong, Steve Cassidy, and Pam Peters. “Updating the ICE annotation system: Tagging, parsing and validation”. In: Corpora 6.2 (2011), pp. 115–144.

The creation and validation of the Construction Complexity Calculator

Chris Cooper

Rikkyo University

In recent years there has been a growing interest in the use of complexity measures, which represent theoretical perspectives of language acquisition, to evaluate the progression of L2 users' proficiency (Biber et al., 2020; Bulté et al., 2024; Kyle, 2016). A recently suggested measure for the English language (Nelson, 2024) is grounded in complexity theory (Larsen-Freeman, 1997) and construction grammar (Goldberg, 2003), quantitatively representing construction complexity through entropy-based calculations. In this presentation, the process of creating the Construction Complexity Calculator will be described, along with tool validation. The tool is designed to increase accessibility to Nelson's (2024) measure for non-technical users. Python code has also been provided so interested researchers can amend the code to suit a variety of research needs such as use with languages other than English or with different part-of-speech taggers that are more suitable for the texts of interest. The tool can be used with individual texts or whole corpora. To validate the tool, complexity scores for the ICNALE corpus were compared with Nelson's (2024) results and complexity scores were calculated for a new dataset, the CEFR Listening Corpus. Complexity scores generally increased across CEFR levels in both datasets. However, the complexity scores in the current study tended to be higher than the original study due to differences in the sentence splitting approach. The sentence tokenization method in the current study, which utilized Stanza (Qi et al., 2020), was deemed to be more appropriate, and it was concluded that the Construction Complexity Calculator accurately calculates Nelson's measure. It is hoped that the tool will allow researchers to calculate the complexity of constructions at the text level for a wide range of research purposes.

List of references

Biber, D., Gray, B., Staples, S., & Egbert, J. (2020). Investigating grammatical complexity in L2 English writing research: Linguistic description versus predictive measurement. Journal of English for Academic Purposes, 46, 100869. https://doi.org/10.1016/j.jeap.2020.100869

Bulté, B., Housen, A., & Pallotti, G. (2024). Complexity and difficulty in second language acquisition: A theoretical and methodological overview. Language Learning, Advance online publication. https://doi.org/10.1111/lang.12669

Goldberg, A. E. (2003). Constructions: A new theoretical approach to language. Trends in Cognitive Sciences, 7(5), 219–224. https://doi.org/10.1016/S1364-6613(03)00080-9

Kyle, K. (2016). Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication [Doctoral dissertation, Georgia State University]. https://doi.org/10.57709/8501051

Larsen-Freeman, D. (1997). Chaos/complexity science and second language acquisition. Applied Linguistics, 18(2), 141–165. https://doi.org/10.1093/applin/18.2.141

Nelson, R. (2024). Using constructions to measure developmental language complexity. Cognitive Linguistics, 35(4), 481–511. https://doi.org/10.1515/cog-2023-0062

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 101–108. https://doi.org/10.18653/v1/2020.acl-demos.14

Dimensions of register variation in contemporary German

Andressa Costa[1], **Tony Berber Sardinha**[2]

[1]Karlsruhe Institut of Technology (KIT); [2]Pontifical Catholic University of São Paulo (PUC-SP)

There are few quantitative studies that look at how contemporary German is used across different spoken, written, and online registers. At the same time, most descriptive research on German language use aims to analyze the use of individual linguistic characteristics in single registers, focusing largely on such registers as newspaper and magazines, conversation, and academic texts. Previous languages that have been subject to Multidimensional (MD) Analysis include English (Biber, 1988), Spanish (Biber et al., 2006; Parodi, 2007), Portuguese (Berber Sardinha, Kauffmann & Acunzo, 2012, 2014), Somali (Biber, 1995), and Korean (Biber, 1995), among others. Despite German being a significant European language, it has not yet been extensively analyzed using MD Analysis. This paper aims to address this gap by carrying out a corpus-based investigation of register variation in German across a wide range of registers in different discourse domains from an MD perspective (Biber 1988; Berber Sardinha & Veirano Pinto, 2014, 2019). The corpus consists of 3,068 texts from 52 different registers, totaling ca. 14.5 million words. The corpus was tagged for part of speech and parsed with RFTagger, TreeTagger and ParZu. The tagged texts underwent fix-tagging to resolve issues with the annotation, and the resulting tags were post-processed with a specialized tag count program developed for this project. The normed counts of the individual linguistic features were entered in a factorial analysis using R (Egbert & Staples 2019, Murphy 2021). A model with five factors was interpreted to determine the underlying functions, giving rise to the dimensions. The texts were scored on each dimension. The dimensions provide a general description of how German is used in the examined registers. In the paper presentation, we will introduce and illustrate the dimensions with examples. We will also describe the variation across the different registers with respect to the dimensions.

List of references

Berber Sardinha, T., Kauffmann, C., & Acunzo, C. M. (2012). Dimensions of Variation in Brazilian Portuguese. English Department. Northern Arizona University.

Berber Sardinha, T., Kauffmann, C., & Acunzo, C. M. (2014). Dimensions of register variation in Brazilian Portuguese. In T. Berber Sardinha & M. Veirano Pinto (Eds.), Multi-Dimensional Analysis, 25 years on: A Tribute to Douglas Biber (pp. 35-80). Amsterdam/Philadelphia, PA: John Benjamins.

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

Biber, D. (1995). Dimensions of Register Variation - A Cross-Linguistic Comparison. Cambridge: Cambridge University Press.

Biber, D., Davies, M., Jones, J. K., & Tracy-Ventura, N. (2006). Spoken and written register variation in Spanish: A multi-dimensional analysis. Corpora, 1(1), 1-37.

Egbert, J; Staples, S. (2019). Doing Multi-Dimensional Analysis in SPSS, SAS, and R In: Berber Sardinha, T.; Veirano Pinto, M.. Multi-Dimensional Analysis: Research Methods and Current Issues. London; New York: Bloomsbury Academic, p. 125-144

Murphy, P. (2021). Exploratory Factor Analysis. RPubs. Online Resource: http://rpubs.com/pjmurphy/758265

Parodi, G. (2007). Variation across registers in Spanish: Exploring the El-Grial PUCV Corpus. In G. Parodi (Ed.), Working with Spanish Corpora (pp. 11-53). London: Continuum.

Linguistic variation in identity-first versus person-first language among autism community stakeholders in a Reddit corpus

Alyssa Marie Crezee, **Earl Kjar Brown**

Brigham Young University

Presently, there exists competing variation between identity-first language (IFL, e.g., *autistic person*) and person-first language (PFL, e.g., *person with autism*) preference among autism community stakeholders (Dunn & Andrews, 2015). Typically, autistic people prefer IFL and other community stakeholders (e.g., parents and professionals) prefer PFL (Taboas et al., 2023). However, it remains unclear how IFL and PFL usage vary in frequency and form. In the present study, a 13-million-token Autism Reddit Corpus (ARC) was created from 11,489 posts (including comments) across 16 autism-related subreddits. Customized queries were performed in LancsBox X (Brezina & Platt, 2024) to identify virtually all instances of various linguistic realizations of IFL and PFL referencing autistic individuals. Chi-square analysis (p<0.01) with Cramer's V effect sizes were used to identify significant differences in IFL and PFL usage across subreddits, targeting the different stakeholders, and concordance analysis was used to qualitatively explore reasons for IFL/PFL form variation. Variation between IFL and PFL usage across subreddits in ARC reveals several significant differences across subreddit communities. More specifically, the results showed that IFL (1292.05 per million) occurred more frequently overall than PFL (117.43 per million). In addition, PFL occurred more frequently in subreddits most relevant to parents of autistic individuals (two subreddits) than in subreddits most relevant to the individuals themselves (14 subreddits). Variation between the most prototypical form of IFL (*autistic* NOUN) and less frequent forms of IFL (*autistics* and *autist(s)*) across subreddits and posts will be discussed. These findings contrast somewhat with previous survey and forced-choice task data regarding stakeholder preferences (Taboas et al., 2023). Furthermore, the methods in this study can be applied to other communities and platforms to explore IFL and PFL variation as it relates to preference, usage, and identity.

List of references

Brezina, V ., Platt, W. (2024). #LancsBox X 5.0.3 [software package], lancsbox.lancaster.ac.uk

Dunn, D. S., & Andrews, E. E. (2015). Person-first and identity-first language: Developing psychologists' cultural competence using disability language. American Psychologist, 70(3), 255.

Taboas, A., Doepke, K., & Zimmerman, C. (2023). Preferences for identity-first versus person-first language in a US sample of autism stakeholders. Autism, 27(2), 565-570.

Intersecting discourses of climate and migration in research-based news

Niall Curry[1], **Dario Del Fante**[2]

[1]Manchester Metropolitan University; [2]Universita' degli Studi di Ferrara

Garnering public engagement with research has become a central concern in contemporary academia. This centrality has given rise to a growing interest in the social and linguistic features of public-oriented academic texts, including those constituting research-based news. Studies of research-based news demonstrate the capacity for texts, such as academic news blog posts, to reflect both disciplinary- and culturally-situated epistemologies (Curry, 2024) and to shed light on the ideological values underpinning global knowledge construction (Curry & Brookes, 2025). Research in this area has identified cross-cultural variation in the discursive construction of the climate crisis (Curry, 2024; Curry & Brookes, 2025) as well as cross-cultural and intertextual variation in the framing of knowledge of the COVID-19 pandemic (Curry & Pérez-Paredes, 2021; Zou & Hyland, 2024), for example. This paper adds to this growing canon with a focus on the intersecting discourses of climate and migration. Given that both the climate crisis and migration have become central matters of debate and conflict in wider society (Collins & Nerlich, 2016; Del Fante & Taylor, 2024), there is a need for a critical understanding of how the knowledge of these concepts is constructed and framed for the public.

To address this need, this paper presents a modern-diachronic corpus-assisted discourse analysis of climate and migration-themed texts from The Conversation. Drawing on Brookes and Curry (2024), we use keyword analyses to identify stable and shifting discourses across time and space, with a focus on texts produced in Australia, the UK, and the US between 2011 and 2024 (inclusive). The analysis highlights convergences in the construction of climate and migration-themed knowledge and pinpoints divergences in this knowledge-making that are disciplinarily- and culturally-situated. This analysis allows us to unpack the ideological values underpinning knowledge construction and make recommendations for more effective, transparent, and reflective forms of public-oriented research communication.

List of references

Brookes, G., & Curry, N. (2024). The changing discourses on Islamophobia in the UK press: A modern-diachronic corpus-assisted study. Journal of Corpus and Discourse Studies, 7, 101-124. https://doi.org/10.18573/jcads.128

Collins, L. C., & Nerlich, B. (2016). Uncertainty discourses in the context of climate change: A corpus-assisted analysis of UK national newspaper articles. Communications, 41(3), 291-313. https://doi.org/10.1515/commun-2016-0009

Curry, N. (2024). Questioning the climate crisis: A contrastive analysis of parascientific discourses. Nordic Journal of English Studies, 23(2), 235-267. https://doi.org/10.35360/njes.v23i2.39190

Curry, N., & Brookes, G. (2025). The discursive framing of the climate and health polycrisis in English, French and Spanish. In Parnell T, Van Hout T, & Del Fante D. (Eds.), Critical approaches to polycrisis: Discourses of conflict, migration, risk and climate. Palgrave Macmillan.

Curry, N. & Pérez-Paredes, P. (2021). Stance nouns in COVID-19 related blog posts: A contrastive analysis of blog posts published in The Conversation in Spain and the UK. International Journal of Corpus Linguistics, 26(4), 469-497. https://doi.org/10.1075/ijcl.21080.cur

Del Fante, D., & Taylor, C. (2024). Migration discourses in times of crisis. Critical Approaches to Discourse Analysis Across Disciplines, 14(2), i-viii. https://doi.org/10.21827/cadaad.14.2.41616

East vs. West: Conceptual leitmotifs over the 30 years of post-socialist mainstream media

Václav Cvrcek[1], **Masako Fidler**[2]

[1]Charles University; [2]Brown University

This paper explores diachronic change in conceptual associations in the mainstream Czech language press. It tracks two notions, East and West over the 30 years after the fall of communism (1991–2022). The project reflects how Czech media frames the two concepts during this eventful post-communist period as the Czech Republic went through economic transition, joined the NATO and EU, and witnessed the Russian invasion of Ukraine.

Conceptual associations are drawn from words ("associated words") that regularly co-occur with the seed word within each media article. These associations create persistent ideas (leitmotifs) connected with the concept of East and West. The methodologies and the results in this study differ from collocation analysis, which prototypically points to word usage (Baker 2010, Gabrielatos et al 2012, McEnery et al 2019). Associations are identified by quantitative methods using probability of co-occurrence via application of Market Basket Analysis (Han et al. 2011; Information Resources Management Association 2014) or similarity/correlation of frequency trends in time (Companions) (Cvrček and Fidler 2024).

The associated words of the seed words *East* and *West* show a clear shift in where the two concepts occur: the period immediately after the Soviet bloc collapse is marked by a search for identity and the position of Czechia as part of Central Europe (with associated words, e.g., *Central, Central-European*); later, *East* and *West* are embedded in articles featuring US-Russia geopolitics (e.g., *America, Washington, Russia, Soviet*), and finally in articles on Russia and Ukraine (e.g., *Putin, Ukraine, separatist*). While the results relate to the ongoing events, they also inform us that the non-existing USSR, persistently looms over in the media even in 2010s to 2022 (*Soviet, USSR*). It is also noteworthy that the EU has little weight in conceptualizing East and West even during the Czechia's accession to the EU.

List of references

Baker, P. 2011. Times may change, but we will always have money: Diachronic variation in recent British English. Journal of English Linguistics 39(65), 65-88.

Cvrček, V. and M. Fidler. 2024. From news to disinformation: unpacking a parasitic discursive practice of Czech pro-Kremlin media. Scando-Slavica, 70(1). https://www.tandfonline.com/doi/full/10.1080/00806765.2024.2317374

Fidler, M. and V. Cvrček. 2023 Zone-Flooding As a Discursive Strategy of Czech Anti-System News Portals. Journal of Slavic Linguistics, 31(1-2), 61-97, https://ojs.ung.si/index.php/JSL/article/view/248.

Gabrielatos, C., T. McEnery, P. J. Diggle & P. Baker. 2012. The peaks and troughs of corpus-based contextual analysis. International Journal of Corpus Linguistics, 37(2), 151-175.

Han, J., M. Kamber, and J. Pei. 2011. Data Mining: Concepts and Techniques. 3rd edition. Haryana, India – Burlington, MA: Morgan Kaufmann.

Information Resources Management Association. (2014). Computational Linguistics: Concepts, Methodologies, Tools, and Applications (1 edition). IGI Global.

McEnery, T., V. Brezina and H. Baker. 2019. Usage Fluctuation Analysis A new way of analysing shifts in historical discourse. International Journal of Corpus Linguistics 24(4), 413–444.

Conversational Persian: Insights from a learner corpus

Sepideh Daghbandan

University of Edinburgh

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 the Conversational Persian by learners of Persian remains scarce. Therefore, this study aims to explore the use of Conversational Persian by language learners using a learner corpus. To this end, a spoken learner corpus, namely, the Learner of Persian Spoken Corpus (LoPSC) was compiled. LoPSC is the first spoken corpus collected from learners of Persian. Data from LoPSC consists of approximately 40,000 words of transcribed audio recordings from conversations between advanced learners of Persian. After the compilation of LoPSC, 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 of Persian speakers living in Iran. 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 in their use of discourse markers. That is, the learners used significantly fewer discourse markers compared to their L1 speaker counterparts. The two groups of speakers also used different pragmatic functions for the same discourse markers. This study has three main contributions. First, it provides empirical findings in a novel context, namely, the use of Conversational Persian by learners. Second, this study also provides further empirical evidence on how learners use discourse markers in Conversational Persian, especially in comparison with L1 speakers of Persian. Finally, as the first study to compile and analyse a spoken learner corpus in Persian, this study also provides insights into the challenges of compiling a learner corpus in this language, especially regarding the conversational register of Persian

Comparing health care providers' and patients' discourses of chronic pain: A corpus-based discourse analysis

Jana Declercq[1], Mathew Gillings[2], Isolde van Dorst[3]

[1]University of Antwerp; [2]Vienna University of Economics and Business; [3]Vienna University of Economics and Business

Using a CADS approach, this paper compares patients' and health care providers' (HCPs) discourses on illness, the body, and pain in pain clinic consultations. Whilst traditional work in health communication has assumed differences in language are a result of differences in medical knowledge, it is possible that this could also be a matter of fundamentally different sociocultural understandings. This is particularly relevant for chronic pain; usually treated by HCPs using an integrated perspective on mental and physical aspects of pain, patients often consider it as a purely physical, biomedical issue. At the same time, chronically ill patients often have a considerable degree of expertise about their illness, and the binary of a professional medical expert versus patient does thus not adequately reflect the complexity of talking about their illness.

This raises the question whether and how this binary is reflected in discourse in clinical care settings. We study the similarities and differences in the discourses produced by patients and their HCPs, both in their use of jargon, but also in different underlying perspectives on illness, the body, and pain. To do this, we analyse a 232,952-word-corpus of 38 Dutch-language consultations between patients with chronic pain and HCPs, collected at a Belgian pain clinic. We use keyness analysis as a starting point, building on this using other tools such as concordance and collocation analysis. Preliminary results point, among other things, to a difference in thematic focus - HCPs discuss diagnostic and treatment options, medication, and their colleagues, while patients' keywords capture aspects of bodily sensations and experiences, and social relations.

We will further explicate what possible interpretations of these findings are. With this, we hope to contribute to a better understanding of chronic pain in the clinical context and beyond, and, in doing so, support HCPs and patients in care encounters.

Artificial intelligence in songwriting: A lexical multi-dimensional approach

Maria Claudia Delfino, Tony Berber Sardinha

Pontifical Catholic University of Sao Paulo

The rapid rise of Artificial Intelligence (AI) chatbots, such as ChatGPT, has sparked intense debate regarding their capacity to replicate or even replace human-authored texts (Berber Sardinha, 2024). While AI-generated outputs excel in structured tasks like weather forecasting or financial reports, their role in creative domains - particularly songwriting - has attracted growing interest. This study investigates the ability of AI to emulate human creativity in English-language song lyric composition. Using Lexical Multi-Dimensional Analysis (Berber Sardinha and Fitzsimmons-Doolan, 2024), we constructed a purpose-specific corpus of 4,000 song lyrics (1.2 million words) across five musical genres: country, pop, rap, rock and soul. The corpus is evenly divided between human-authored and AI-generated texts, with balanced representation across genres. AI-generated lyrics were sourced from ChatGPT, Google's Gemini, and Meta's Llama (both mainstream and uncensored versions), with each system contributing 25% of the AI subcorpus. Our analysis identifies five key dimensions underlying discourse variation in song lyrics, enabling comparisons by origin (AI vs human), AI model, and musical genre. The discourses unveiled by the dimensions are social justice vs romance, reality vs transcendence, rural vs urban, individualism vs collectivism and extroversion, physicality vs introversion, emotions. Distinctively, this study incorporates discourse-level features containing lexical elements to differentiate between human and AI-generated lyrics. The findings reveal significant patterns of convergence and divergence between human and AI-generated lyrics, shedding light on the creative potential and limitations of AI in this culturally and artistically significant domain. Additionally, the study explores how AI-generated discourse reflects underlying ideologies, raising broader implications for the evolution of AI creativity and its influence on cultural production.

Acknowledgement

The authors acknowledge the financial support of the following organizations: São Paulo Research Foundation (FAPESP), Grant #2022/05848-7; National Council for Scientific and Technological Development (CNPq), Grants # 310140/2021-8, 403130/2024-7, 444019/2024-3.

List of references

Berber Sardinha, T., & Fitzsimmons-Doolan, S. (in press). Lexical Multidimensional Analysis: Identifying Discourses and Ideologies. Cambridge: Cambridge University Press.

Berber Sardinha, T. (2024). AI-generated vs human-authored texts: A Multidimensional comparison. Applied Corpus Linguistics, 4(1), 100083. https://doi.org/https://doi.org/10.1016/j.acorp.2023.100083

Delfino, M. C. N., Berber Sardinha, T., & Collentine, J. G. (2023). Dimensões de variação lexical e acústica na música popular em inglês: um estudo baseado em corpus [Lexical and acoustic dimensions of variation in popular music in English: A corpus-based study]. Cadernos de Estudos Linguísticos, 65, e023025. https://periodicos.sbu.unicamp.br/ojs/index.php/cel/article/view/8671801/33021

Searching for criterial features to characterise L2 accuracy profiles in two task types by Spanish learners of English at CEFR B1 level: Insights from the FineDesc learner corpus

María Belén Díez-Bedmar[1], Jennifer Thewissen[2]

[1]University of Jaén; [2]Universiteit Antwerpen

Accuracy, one of the Complexity, Accuracy, and Fluency (CAF) measures, has been a focal point of Learner Corpus Research. While numerous studies have examined how accuracy varies with proficiency (e.g., Díez-Bedmar, 2011, 2015, 2018; Gráf & Huang, 2022; Hawkins & Filipović, 2012; Thewissen, 2015), the influence of task type on L2 accuracy profiles has received comparatively little attention (see exceptions: Alexopoulou et al., 2017; Lan, 2015; Lyashevskaya et al., 2022). This gap has hindered the identification of criterial features (Hawkins & Filipović, 2012) that distinguish error types across different task types within the same proficiency level.

This study investigates criterial features in two task types - emails and narrative texts -produced by Spanish learners of English at CEFR B1 level, using data from the FineDesc Learner Corpus. The dataset consists of 100 emails and 100 narrative texts written by the same 100 students. Errors were manually tagged by two experienced annotators following the Louvain Error Tagging Manual. Non-parametric statistical tests were run, and effect sizes were also examined for non-significant results (Plonsky, 2015).

The results reveal that task type significantly impacts accuracy profiles. Narrative writing triggered a greater variety and frequency of errors, particularly in the verb phrase domain. These findings are interpreted in light of the Task-Based Language Teaching (TBLT) framework (e.g., Skehan & Foster, 2012). Additionally, task type was also found to affect annotator agreement, suggesting task-specific challenges in error analysis.

This study highlights the critical role of task type in shaping L2 accuracy profiles and provides methodological insights for error annotation. By identifying criterial features across task types, the research offers valuable implications for Learner Corpus Research, assessment practices, and task design in language teaching. The findings regarding the task type effect on error annotation also provide insights into the issues in the annotation process (Díez-Bedmar, 2020).

List of references

Alexopoulou, T., Michel, M., Murakami, A., & Meurers, D. (2017). Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques. Language Learning, 67(1), 180–208. https://doi.org/10.1111/lang.12232

Díez-Bedmar, M. B. (2011). Spanish pre-university students' use of English: CEA results from the University Entrance Examination. International Journal of English Studies, 11(2), 141-158. https://doi.org/10.6018/ijes/2011/2/149681

Díez-Bedmar, M. B. (2015). Article use and criterial features in Spanish EFL writing: A pilot study from CEFR Benjamins. A2 to B2 levels. In M. Callies & S. Götz (Eds.), Learner corpora in language testing and assessment (pp. 163–190). John Benjamins.

Díez-Bedmar, M. B. (2018). Fine-tuning descriptors for CEFR B1 level: insights from learner corpora. ELT Journal, 72(2), 199–209. https://doi.org/10.1093/elt/ccx052

Gráf, T., & Huang, L. (2022). Persistent errors in spoken English among Taiwanese and Czech learners at CEFR B2 and C1. In A. Leńko-Szymanska & S. Götz (Eds.), Complexity, Accuracy and Fluency in Learner Corpus Research (pp. 137–158). John Benjamins.

Hawkins, J. A., & Filipović, L. (2012). Criterial features in L2 English: Specifying the reference levels of the Common European Framework. Cambridge University Press.

Lan, N.T. (2015). The Effect of Task Type on Accuracy and Complexity in IELTS Academic Writing. VNU Journal of Science, 45-63.

Lyashevskaya, O., Vinogradova, O., & Scherbakova, A. (2022). Accuracy, syntactic complexity and task type at play in examination writing. In A. Leńko-Szymańska & S. Götz (Eds.). Complexity, accuracy and fluency in learner corpus research (pp. 241-272). John Benjamins.

Plonsky, L. (2015). Statistical power, p values, descriptive statistics, and effect sizes: A “Back-to-Basics” Approach to Advancing Quantitative Methods in L2 Research. In L. Plonsky (Ed.). Advancing quantitative methods in second language research (pp. 23-45). Routledge.

Thewissen, J. (2015). Accuracy across proficiency levels: A learner corpus approach. Presses Universitaires de Louvain.

Lexical diversity profile In L2 English learner writing across genres

Thi Kieu Trinh Dinh, Philip Durrant

University of Exeter

Lexical diversity across genres in L2 learner writing has yielded conflicting results. While some studies have found a significant difference in the diversity of vocabulary use in different genres (Taylor, 2024; Yoon & Polio, 2016), others have not (Abdi et al., 2024; Lee, 2021). Moreover, there has been little consideration of why some genres are found to have more diverse vocabulary than others. Given that genre awareness is an important competence that develops throughout the course of L2 learning at school, this paper aims to examine if and why there is a significant difference in the diversity of vocabulary in two genres (inarrative and opinion), across different year groups.

A total of 604 texts responding to 95 distinct tasks written by 120 young L2 learners from Year 8 to Year 11 in Norway was examined using quantitative measures of diversity. Narrative writing was found to have significantly more diverse vocabulary than opinion writing, as measured by Measure of Textual Lexical Diversity (McCarthy, 2005). By calculating the mean distance between tokens of types that are repeated more than once, we also found that there is less distance between repetitions in opinion than in narrative writing. Moreover, to examine if more lexically diverse narrative writing means more types of different meanings were used, latent semantic analysis was employed to measure the mean semantic similarity of all lexical types in a text based on the ukWaC semantic space (Baroni et al., 2009). No significant difference was found between two genres but across year groups, in which older learners used significantly less semantically related types in the opinion genre. Writing tasks were also found to be a significant effect in these findings. This presentation will discuss why lexical use in different tasks could result in such differences and propose pedagogical implications regarding genre awareness.

List of references

Abdi T., M., Lee, J., & Wang, Y. (2024). Capturing linguistic features of writing in two genres over time. Reading and Writing, 37(3), 787-809.

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. Language resources and evaluation, 43, 209-226.

Lee, J. (2021). Using corpus analysis to extend experimental research: Genre effects in L2 writing. System, 100, 102563.

McCarthy, P. M. (2005). An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD) (Doctoral dissertation, The University of Memphis).

Taylor, C. (2024). Exploring the Impact of Genre on Syntactic and Lexical Complexity in L2 Written English. Master Thesis. University of Gävle.

Yoon, H.-J., & Polio, C. (2016). The Linguistic Development of Students of English as a Second Language in Two Written Genres. TESOL Quarterly, 51(2), 275–301.

Decoding thinking patterns of different psychological states: A machine learning study on textual syntax

Xiaowei Du[1], **Yaqian Shi**[2]

[1]Dalian Maritime University; [2]Huazhong University of Science and Technology

Syntactic features serve as critical linguistic markers of psychological states, with structural patterns in language production revealing clinically relevant insights. While existing research has been constrained by simplified measures and traditional statistical methods, this study introduces an advanced approach that combines comprehensive quantitative syntactic analysis with interpretable machine learning techniques. We systematically analyzed 16 syntactic complexity indices extracted from forum posts across three groups: individuals with depression, individuals with bipolar disorder, and healthy controls. Our machine learning classification framework demonstrated that syntactic complexity metrics exhibit strong discriminative capability, confirming syntactic patterns as robust psycholinguistic indicators. More importantly, normalized dependency distance (NDD) emerged as the most predictive indicator for text classification across psychological states, with mental disorder groups showing markedly distinct patterns compared to controls. These findings have significant clinical implications for mental health assessment and intervention, as well as important methodological implications for syntactic research.

List of references

Axiotis, K., Abu-al-haija, S., Chen, L., Fahrbach, M., & Fu, G. (2023). Greedy PIG: Adaptive Integrated Gradients (arXiv:2311.06192). arXiv. http://arxiv.org/abs/2311.06192

Björklund, J., & Zechner, N. (2017). Syntactic methods for topic-independent authorship attribution. Natural Language Engineering, 23(5), 789–806. https://doi.org/10.1017/S1351324917000249

Casal, J. E., & Lee, J. J. (2019). Syntactic complexity and writing quality in assessed first-year L2 writing. Journal of Second Language Writing, 44, 51–62. https://doi.org/10.1016/j.jslw.2019.03.005

Chen, X., Alexopoulou, T., & Tsimpli, I. (2021). Automatic extraction of subordinate clauses and its application in second language acquisition research. Behavior Research Methods, 53(2), 803–817. https://doi.org/10.3758/s13428-020-01456-7

Crossley, S. A., & McNamara, D. S. (2014). Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. Journal of Second Language Writing, 26, 66–79. https://doi.org/10.1016/j.jslw.2014.09.006

Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., Jha, S., Agrawal, M., Dziurzynski, L. A., Sap, M., Weeg, C., Larson, E. E., Ungar, L. H., & Seligman, M. E. P. (2015). Psychological Language on Twitter Predicts County-Level Heart Disease Mortality. Psychological Science, 26(2), 159–169. https://doi.org/10.1177/0956797614557867

Flekova, L., Preoţiuc-Pietro, D., & Ungar, L. (2016). Exploring Stylistic Variation with Age and Income on Twitter. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 313–319. https://doi.org/10.18653/v1/P16-2051

Kim, M., & Crossley, S. A. (2018). Modeling second language writing quality: A structural equation investigation of lexical, syntactic, and cohesive features in source-based and independent writing. Assessing Writing, 37, 39–56. https://doi.org/10.1016/j.asw.2018.03.002

Kyle, K. (2016). Measuring Syntactic Development in L2 Writing: Fine Grained Indices of Syntactic Complexity and Usage-Based Indices of Syntactic Sophistication. 202.

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

Lei, L., & Shi, Y. (2023). Syntactic complexity in adapted extracurricular reading materials. System, 113, 103002. https://doi.org/10.1016/j.system.2023.103002

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2018). Feature Selection: A Data Perspective. ACM Computing Surveys, 50(6), 1–45. https://doi.org/10.1145/3136625

Lundberg, S., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions (arXiv:1705.07874). arXiv. http://arxiv.org/abs/1705.07874

Morales, M. R., & Levitan, R. (2016). Speech vs. text: A comparative analysis of features for depression detection systems. 2016 IEEE Spoken Language Technology Workshop (SLT), 136–143. https://doi.org/10.1109/SLT.2016.7846256

Neary-Sundquist, C. A. (2017). Syntactic complexity at multiple proficiency levels of L2 German speech: Complexity in L2 German. International Journal of Applied Linguistics, 27(1), 242–262. https://doi.org/10.1111/ijal.12128

Noecker, J., Ryan, M., & Juola, P. (2013). Psychological profiling through textual analysis. Literary and Linguistic Computing, 28(3), 382–387. https://doi.org/10.1093/llc/fqs070

Zinken, J., Zinken, K., Wilson, J. C., Butler, L., & Skinner, T. (2010). Analysis of syntax and word use to predict successful participation in guided self-help for anxiety and depression. Psychiatry Research, 179(2), 181–186. https://doi.org/10.1016/j.psychres.2010.04.011

Reading concordances at scale: Body parts in English and German 19th-century fiction

Nathan Dykes, Stephanie Evert, Michaela Mahlberg, Alexander Piperski

Friedrich-Alexander-Universität Erlangen-Nürnberg

Concordance reading is central to corpus linguistics. It helps to identify recurring patterns while considering a substantial amount of co-text. An essential part of finding such repeated linguistic forms is re-arranging concordance lines. This is typically done via sorting, where the sequence of lines is determined by a fixed criterion specified by the researcher. Most concordance tools offer alphabetical sorting on a specific position left or right of the node. While this is a popular technique, it has its limitations, especially in identifying semantic similarities between lines with different wordings. This paper presents a novel approach to concordance reading by clustering concordance lines based on their overall similarity within an adjustable context window. We compare three clustering algorithms using our dedicated Python library (TF-IDF, SpaCy word embeddings, Sentence Transformers). Our case study compares body part nouns in corpora of German and English 19th-century novels (Mahlberg, 2016; Evert et al., 2024), which are available through the web application CLiC (Mahlberg et al., 2020). In an approach similar to Mahlberg et al. (2020), we examine linguistic patterns associated with various nouns (e.g. mouth, foot, cheek and their German translation equivalents). We investigate similarities and differences regarding how body parts are grouped together by different algorithms in both languages. The first comparison step concerns the variety and distribution of clusters. For German, we find a smaller variety of frequently mentioned body part nouns. In both languages, while some clusters contain specific body part nouns, e.g. foot, other clusters group together nouns used in similar contexts, e.g. interactions between two characters. Our comparison of algorithms shows that Sentence Transformers are most useful to identify patterns that can be interpreted in terms of narrative functions of fiction. Based on these insights, we suggest applications in the analysis of fiction.

List of references

Evert, S., Finlayson, N., Mahlberg, M., & Piperski, A. (2024). DE19: Deutsche Romane des 19. Jahrhunderts. Reading Concordances in the 21st Century.

Mahlberg, M. (2016). 19th Century Reference Corpus. Centre for Corpus Research.

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

Mahlberg, M., Wiegand, V., & Hennessey, A. (2020). Eye language – body part collocations and textual contexts in the nineteenth-century novel. In L. Fesenmeier & I. Novakova (Eds.), Phraséologie et stylistique de la langue littéraire / Phraseology and Stylistics of Literary Language. Approches interdisciplinaires / Interdisciplinary Approaches (pp. 143–176). Peter Lang.

Continuous keyness? Identifying words and features that correspond with continuous extralinguistic variables

Jesse Egbert, Elizabeth Hanks, Doug Biber

Northern Arizona University

For decades, keyword analysis has been a fixture in corpus linguistics and discourse analysis (see Scott, 1997). More recently, key feature analysis has been introduced to apply keyness to linguistic features (Egbert & Biber, 2023). Keyword and key feature analysis provide researchers with a list of linguistic items that are more prevalent in one (target) corpus when compared with another (reference) corpus. A limitation of these methods is that they only allow comparisons between two groups. This works well when one is interested in a variable that is inherently dichotomous, like two dialects (e.g. Baker, 2017) or one register versus all others (e.g. Biber & Egbert, 2018). But what if we want to know which words and linguistic features are most strongly associated with a *continuous* extralinguistic variable like age, time, or the degree to which texts fulfill a communicative purpose? One approach is to discretize a continuous variable into a dichotomous variable—as Baker & Brookes (2022) did with age and Hoffman et al. (2020) did with time—and then apply traditional keyword analysis. However, this eliminates valuable (co)variance data by reducing truly quantitative scales down to two categories. It also forces researchers to contrive arbitrary cut-off points to discretize the quantitative scale into two categories. To address these challenges, we propose new methods that allow us to extend the categorical notion of *keyness* to the continuous notion of *correspondence* between continuous extralinguistic variables and linguistic items: Corresponding Word Analysis (similar to Grieve et al., 2017) and Corresponding Feature Analysis. We also propose appropriate adaptations for non-prevalent linguistic items and small texts. We then demonstrate how these methods can be successfully applied to identify words and features that correspond to continuous communicative purposes in British conversation and the degree to which texts can be referred to as conversations.

List of references

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

Baker, P., & Brookes, G. (2022). Analysing language, sex and age in a corpus of patient feedback: A comparison of approaches. Cambridge University Press.

Biber, D. & Egbert, J. (2018). Register variation online. Cambridge University Press.

Egbert, J. & Biber, D. (2023). Key feature analysis: a simple, yet powerful method for comparing text varieties. Corpora, 18(1), 121-133.

Grieve, J., Nini, A., & Guo, D. (2017). Analyzing lexical emergence in Modern American English online. English Language & Linguistics, 21(1), 99-127.

Hofmann, K., Marakasova, A., Baumann, A., Neidhardt, J., & Wissik, T. (2020, May). Comparing lexical usage in political discourse across diachronic corpora. In Proceedings of the Second ParlaCLARIN Workshop (pp. 58-65).

Scott, M. (1997). PC analysis of key words—and key key words. System, 25(2), 233-245.

How we entangled the "voices of the enslaved", and what you could research through it

Irene Elmerot[1], Leif-Jöran Olsson[2], Klas Rönnbäck[2]

[1]Stockholm University; [2]University of Gothenburg

The most extensively researched aspect of the history of slavery is arguably that of enslaved individuals in North America. Their personal accounts have been largely disregarded, but their tales were written down in a few projects, such as the *Federal Writers' Project* (Library of Congress 2001) and *Documenting the American South* (Andrews [no date]). Our comprehensive, open access, annotated corpus of these texts offers a unique opportunity to explore different facets of the lives of enslaved individuals, thereby contributing to the advancement of knowledge about this period of history. The texts includes both standardised and colloquial (African American English) transcripts, and are extracted from PDF files of the original books – all issues which made the lemmatisation and overall annotation of the corpus very complex.

In social science research, living standards among diverse populations are a main subject. However, enslaved individuals have been largely excluded from these studies, due to the absence of crucial data for understanding living conditions. Historians have only employed samples of the above-mentioned projects to study specific aspects of living standards.

To address these challenges and move our knowledge forward, we digitise the PDF files anew, for enrichment and annotation. Using the *Sparv* tool (Hammarstedt et al. 2022), correcting and re-reading the texts, we trained the tool sufficiently for research purposes. The corpus can be utilised in social science and huanities research, allowing for the analysis of how formerly enslaved individuals perceived their lives before, during and after slavery. For example, we perform named entity recognition analyses to create networks of the geographical differences and the differences between individual interviewers and editors.

The presentation will show original (PDF format) texts to address issues in corpus creation and will also introduce an initial version of the corpus and exemplify analyses to be conducted at a subsequent stage.

List of references

References

Andrews, W.L. [no date]. Scholarly Bibliography of Slave and Ex-Slave Narratives. Documenting the American South. Available at: https://docsouth.unc.edu/neh/biblintro.html.

Hammarstedt, M., Schumacher, A., Borin, L. and Forsberg, M. 2022. Sparv 5 User Manual. Gothenburg: Språkbanken Text. Available at: https://spraakbanken.gu.se/en/tools/sparv.

LibraryofCongress. 2001. Collection: Born in Slavery: Slave Narratives from the Federal Writers' Project, 1936 to 1938. Available at: https://www.loc.gov/collections/slave-narratives-from-the-federal-writers-project-1936-to-1938/about-this-collection/.

Writing like someone else: A corpus of human and AI imitative language

Mel Evans

University of Leeds

Corpus approaches to idiolectal style have provided insights at theoretical, methodological and applied levels (e.g. Burrows and Craig 2012, Evans 2018, Wright 2017, Nini 2018, 2023). Yet, there is still much to investigate in terms of how humans comprehend individual style, as well as what defines individual style at all. With generative AI, the ability of not only profiling, but successfully imitating a writer's style or a performer's voice, living or dead and at scale in creative literary contexts is now possible (e.g. the forthcoming Michael Parkinson podcast). Whilst a commercial challenge to human creative practitioners, this also affords an opportunity to better understand idiolectal style, with potential forensic, legal and creative applications.

This paper presents the approach and key findings of a pilot project to build and analyse a corpus of imitative language. "Generated" by human creative writers and AI from shared prompts, the corpus comprises samples of different genres produced in conscious imitation of a literary author. The paper will first discuss the procedure of data creation, collection and preparation, and the ethical considerations involved in generating imitative writing using human participants and AI. The paper will then present findings comparing the linguistic make-up of the human and AI texts, drawing on analytic techniques developed in corpus linguistics and computational stylistics, designed to profile authorial, generic and temporal linguistic signals in literary writing. The results demonstrate that corpus linguistic approaches to linguistic creativity can provide insights into *how* linguistic imitation is undertaken – similarly and differently - by human language users and AI, particularly when focussing on the distribution of keywords, keyword collocations, and patterns of syntactic repetition, especially pronouns, between human and machine.

List of references

References:

Burrows, J. and Craig, H. 2012. Authors and Characters. English Studies. 93(3), pp.292–309.

Evans, M. 2018. Style and chronology: A stylometric investigation of Aphra Behn's dramatic style and the dating of The Young King. Language and Literature. 27(2), pp.103–132.

Nini, A. 2018. An authorship analysis of the Jack the Ripper letters. Digital Scholarship in the Humanities. 33(3), pp.621–636.

Nini, A. 2023. A Theory of Linguistic Individuality for Authorship Analysis. Cambridge University Press.

Wright, D. 2017. Using word n-grams to identify authors and idiolects: A corpus approach to a forensic linguistic problem. International Journal of Corpus Linguistics. 22(2), pp.212–241.

Bootstrapping in corpus linguistics: A practical guide

Stephanie Evert

Friedrich-Alexander-Universität Erlangen-Nürnberg

Many quantitative techniques in corpus linguistics – such as frequency comparisons, keyword and collocation analysis, or measures of productivity and lexical richness – are strongly affected by the fact that corpora aren't random samples of tokens (as most hypothesis tests assume) but rather samples consisting of entire texts or text fragments (e.g. Evert 2006; Gries 2024: 274), leading to increased sampling variation. Recent methodological work in corpus linguistics therefore strongly recommends the use of bootstrapping approaches (Efron & Tibshirani 1993) that estimate the sampling distribution empirically by resampling corpora with replacement at the level of entire texts (Lijffijt et al. 2016; Gries 2022). However, such techniques are rarely applied in practice:

1. none of the standard corpus software tools implement bootstrapped frequency comparisons or keyword and collocation analysis, so researchers are left to their own devices and programming skills;
2. expositions of the methodology in research papers are often very technical, making it difficult for readers to understand exactly what data and algorithms are required;
3. the appropriate interpretation of bootstrapped sampling distributions remains unclear; and
4. bootstrapping algorithms are computationally expensive, especially when applied to large data sets, so substantial programming experience is needed to carry them out with sufficient efficiency (cf. Gries 2024: 275–297).

The purpose of this contribution is to provide a practical guide to bootstrapped frequency analysis in corpus linguistics, covering all the challenges addressed above: (i) how to obtain suitable frequency data from standard corpus software tools; (ii) how to perform bootstrapping efficiently with relatively simple R code; and (iii) how to make sense of bootstrapping results and integrate them e.g. with keyness or collocation measures. Techniques will be illustrated on several use cases: frequency comparison, keyword/collocation analysis, productivity measures, and multivariate analysis. Thoroughly documented and reproducible code examples will be made available as an online supplement.

List of references

Efron, B. & Tibshirani, R. (1993). An Introduction to the Bootstrap. Number 57 in Monographs on Statistics & Applied Probability. Chapman & Hall, CRC, Boca Raton.

Evert, S. (2006). How random is a corpus? the library metaphor. Zeitschrift für Anglistik und Amerikanistik, 54(2):177–190.

Gries, S. T. (2022). Toward more careful corpus statistics: uncertainty estimates for frequencies, dispersions, association measures, and more. Research Methods in Applied Linguistics, 1(1):100002.

Gries, S. T. (2024). Frequency, Dispersion, Association, and Keyness. Number 115 in Studies in Corpus Linguistics. John Benjamins, Amsterdam, Philadelphia.

Lijffijt, J., Nevalainen, T., Säily, T., Papapetrou, P., Puolamäki, K., & Mannila, H. (2016). Significance testing of word frequencies in corpora. Digital Scholarship in the Humanities, 31(2):374–397.

Finding structure in multivariate data

Stephanie Evert[1], Stella Neumann[2], Gerold Schneider[3], Florian Frenken[2]

[1]Friedrich-Alexander-Universität Erlangen-Nürnberg; [2]RWTH Aachen; [3]Universität Zürich

Multivariate analysis is a well-established technique for the corpus-linguistic study of linguistic variation. Here, multidimensional analysis (MDA; Biber 1988) has been particularly influential. It builds on unsupervised factor analysis, which determines latent dimensions of variation from statistical correlations between a large set of lexico-grammatical features. Evert & Neumann (2017) argue that MDA only accounts for major patterns of variation. It cannot detect more fine-grained patterns nor be targeted to specific aspects. Their geometric multivariate analysis (GMA) therefore introduces a minimally supervised intervention to obtain a low-dimensional focus space representing the desired aspects of variation. This is achieved with supervised linear discriminant analysis (LDA, Fisher 1936), which aims to separate proxy categories carefully chosen to guide the multivariate analysis without pre-empting the research question (mathematically, LDA minimises within-group variability while maximising differences between groups). Neumann & Evert (2021) use text categories from the International Corpus of English (ICE; Greenbaum 1996) as their proxy to create a latent "register space", which enables them to study register divergences across varieties of English and uncover subtle differences that would otherwise go unnoticed. However, in dealing with more than one aspect of variation, namely register *and* variety, their study also reveals potential issues of the method: (i) imbalance in the proxy categories, (ii) the combination of LDA dimensions for multiple variables (register and variety), and (iii) confounding factors in the proxy categories (as LDA across all texts aims to minimise differences between texts from multiple varieties in the same category). We present experiments to examine these problems using Neumann & Evert's (2021) data available in their online supplement. We propose a solution via mathematical adaptations of the LDA algorithm and compare them to the original LDA analysis. Results promise a multivariate procedure that can be applied to the complex interplay of multiple sources of variation.

List of references

Biber, D. (1988). Variation Across Speech and Writing. Cambridge University Press, Cambridge.

Evert, S. & Neumann, S. (2017). The impact of translation direction on characteristics of translated texts. A multivariate analysis for English and German. In De Sutter, G., Lefer, M.-A., and Delaere, I., editors, Empirical Translation Studies. New Theoretical and Methodological Traditions, TiLSM 300, pp. 47–80. Mouton de Gruyter, Berlin. https://www.stephanie-evert.de/PUB/EvertNeumann2017/

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7(2), 179–188.

Greenbaum, S. (1996). Comparing English Worldwide: The International Corpus of English. Clarendon Press.

Neumann, S. & Evert, S. (2021). A register variation perspective on varieties of English. In Seoane, E. and Biber, D., editors, Corpus based approaches to register variation, chapter 6, pp. 143–178. John Benjamins, Amsterdam. https://www.stephanie-evert.de/PUB/NeumannEvert2021/

LLMs are here. Is human linguistic annotation dead or evolving?

Javier Fernández Cruz[1], **Carla Fernández Melendres**[2]

[1]Universidad de Málaga; [2]Universidad de Málaga

Textual annotation is crucial in linguistic research as it enriches corpora with linguistic information, broadening research possibilities and enhancing their utility for analysis. However, it is a time-consuming and costly process, especially when requiring specialised domain expertise or annotating a large sample (Ide & Pustejovsky, 2017). This study aims to achieve two objectives: (i) to review the existing annotation methodologies, and (ii) to contribute to the ongoing efforts to establish standards for LLM-based linguistic annotation. To this end, we analysed a corpus of 150 annotated opinion texts (op-eds and tourism reviews).

The emergence of Large Language Models (LLMs) has sparked research on their potential for annotation tasks (Tan et al., 2024). Their ease of use, accuracy, and affordability have fuelled their widespread adoption, marking a paradigm shift in corpus annotation. Still, their rapid growth presents challenges, particularly the lack of established standards, which can lead to low-quality research and invalid results.

Undoubtedly, LLMs offer speed and cost advantages, but their performance varies depending on the task, context, and prompting technique (Alizadeh et al., 2023). While they excel in tasks like sentiment analysis, human annotations often outperform LLMs in complex tasks due to better context understanding and handling of ambiguity (Curry et al., 2024; Törnberg, 2024a; Yu et al., 2024). Moreover, LLMs are prone to biases and misunderstandings, which can lead to unreliable results, underscoring the need for structured methodologies (Törnberg, 2024b).

Overall, a hybrid approach combining human/LLM annotations is recommended to improve accuracy and comprehensiveness. Additionally, implementing a set of standards, e.g., careful model selection, prompt engineering, and rigorous validation, can significantly enhance the reliability of LLM-based annotation. Still, while LLMs offer substantial efficiency gains, their use in linguistic annotation is complex and requires ongoing refinement of methodologies available to linguists. Finally, some advice on making annotation user-friendly are offered.

List of references

Alizadeh, M., Kubli, M., Samei, Z., Dehghani, S., Zahedivafa, M., Bermeo, J. D., Korobeynikova, M., & Gilardi, F. (2024). Open-Source LLMs for Text Annotation: A Practical Guide for Model Setting and Fine-Tuning. https://doi.org/10.48550/arXiv.2307.02179

Curry, N., Baker, P., & Brookes, G. (2024). Generative AI for corpus approaches to discourse studies: A critical evaluation of ChatGPT. Applied Corpus Linguistics, 4(1), 100082. https://doi.org/10.1016/j.acorp.2023.100082

Ide, N., & Pustejovsky, J. (Eds.). (2017). Handbook of Linguistic Annotation. Springer Netherlands. https://doi.org/10.1007/978-94-024-0881-2

Tan, Z., Li, D., Wang, S., Beigi, A., Jiang, B., Bhattacharjee, A., Karami, M., Li, J., Cheng, L., & Liu, H. (2024). Large Language Models for Data Annotation: A Survey. https://doi.org/10.48550/arXiv.2402.13446

Törnberg, P. (2024a). Best Practices for Text Annotation with Large Language Models. https://doi.org/10.48550/arXiv.2402.05129

Törnberg, P. (2024b). Large Language Models Outperform Expert Coders and Supervised Classifiers at Annotating Political Social Media Messages. Social Science Computer Review, 0(0). https://doi.org/10.1177/08944393241286471

Yu, D., Li, L., Su, H., & Fuoli, M. (2024). Assessing the potential of LLM-assisted annotation for corpus-based pragmatics and discourse analysis: The case of apologies. https://doi.org/10.48550/arXiv.2305.08339

How tolerant of errors is AI? Using a learner corpus to explore human-AI interactions

Adriano Ferraresi, **Daniele Polizzi**

University of Bologna

How tolerant of errors is AI? Using a learner corpus to explore human-AI interactions

Current research on the interplay between corpus linguistics and generative AI typically gravitates toward testing the instrumental value of the technology for the discipline, as LLMs begin to populate toolkits (Anthony 2023) and mimic established approaches to language analysis (Curry et al., 2024). The present contribution delves into a less explored area, showing how corpus methods can inform the investigation of AI-generated language, notably in the context of dialogue-based Computer-Assisted Language Learning. Multiple studies advocate for the use of conversational AI in the language classroom (Bibauw et al., 2022), and yet direct observation of students' interactions with these tools is lacking (Han 2024). To tackle these gaps, we present a novel learner corpus design intended to explore emerging features of human-machine written interaction data. Our corpus consists of 326 interactions (722,537 tokens) by as many Italian university students aged 19-25 , with diverse proficiency levels (mostly low-to-upper-intermediate) and including learners with disabilities and learning disorders, to favour equal access to learning opportunities (CAST 2018). The interactions were collected based on a protocol involving two different LLM-based chatbots (ChatGPT and Pi.AI) and two EFL learning scenarios (small talk and roleplay). In addition to introducing the corpus annotation scheme, we present a case study investigating both sides of learner-chatbot interactions. First, we provide quantitative and qualitative evidence of learners' errors in open-ended and task-oriented conversations, annotated following an adapted version of the Louvain Error Tagging Manual (Granger et al., 2022) that includes new tags tailored to digital communication. Second, following up on Cervini and Paone's (2024) classification of intercomprehension strategies, we leverage corpus methods to evaluate LLMs' responses to those errors, with an eye towards analysing the interaction mechanisms of Generative AI and its potential for language development.

List of references

Anthony, L. (2023). Corpus AI: Integrating Large Language Models (LLMs) into a Corpus Analysis Toolkit. Presentation given at the 49th Annual Conference of the Japan Association for English Corpus Studies (JAECS), Kansai University, Osaka, Japan. Available at https://osf.io/srtyd/.

Bibauw, S., W. Van Den Noortgate, T. François and P. Desmet (2022). "Dialogue systems for language learning: a meta-analysis". Language Learning & Technology, 26(1).

CAST. 2018. Universal Design for Learning guidelines. http://udlguidelines.cast.org

Cervini, C., & Paone, E. (2024). Comunicare all'università: quando l'interazione orale si fa plurilingue. Italiano LinguaDue, 16(2), 496-523.

Curry, N., Baker, P., & Brookes, G. (2024). Generative AI for corpus approaches to discourse studies: A critical evaluation of ChatGPT. Applied Corpus Linguistics, 4(1), 100082.

Granger, S., Swallow, H., & Thewissen, J. (2022). The Louvain error tagging manual. Version 2.0.

Han, Z. (2024). "Chatgpt in and for second language acquisition: a call for systematic research". Studies in Second Language Acquisition, 46(2), 301-306.

A data-driven approach to the development of academic medical English at the university of Verona

Carlotta Fiammenghi, **Sharon Hartle**

University of Verona

Academic writing in English poses significant challenges for both non-native and native speakers, who need to master the conventions of academic language alongside specialised professional language [1] [2]. The goal of this collaborative research project between the Department of Foreign Languages and the Department of Medicine at the University of Verona is to create a learner parallel corpus, to investigate the linguistic challenges encountered by advanced learners of academic medical English, and to provide them with tools to assist them in their writing process [3].

The project includes this corpus-based study of academic writing and an EMI course tailored to meet the needs of the Department of Medicine. In our study, first, medical researchers voluntarily provide drafts of their research papers, which are then analysed by researchers from the Department of Foreign Languages. The Markin software is used (Version 4.3.1.1) to annotate errors, signalling error type without providing overt correction; instances of good language use are also highlighted. This is followed by a meeting where researchers discuss potential solutions following a think-aloud protocol [4]. Revised drafts are then re-analysed, and the results are used to create an annotated learner corpus [5]. Finally, a parallel corpus is created using examples of academic papers from top medical journals, chosen by the medical participants based on their publication goals.

Findings suggest that common issues include collocations and word choice, cohesion, and coherence. The collaborative feedback process appears to help participants gain a better understanding of their mistakes.

The research contributes to an understanding of the challenges faced by medical professionals writing in English and offers insights into effective methods for improving academic writing. By creating a learner corpus, the project also provides a resource for both future tutoring and independent learning, ultimately enhancing the ability of medical researchers to publish in international journals.

List of references

[1] Swales, J. M. (2004). Research genres: Exploration and application. Cambridge, UK: Cambridge University Press

[2] Hartle, S. & Cavalieri, S. (2024). Exploring pedagogical approaches in EAP teaching. In Morrison, B. R., MacDiarmid, C., Williams, A. & Haghi, I. (eds.). Proceedings of the 2021 BALEAP Conference (pp. 209-216). Garnet Publishing.

[3] Flowerdew, J. (2017). Corpus-based approaches to language description for specialized academic writing. Language Teaching, 50(1), 90–106. doi:10.1017/S0261444814000378

[4] Zhang, L. J., & Zhang, D. (2019). Think-aloud protocols. In The Routledge handbook of research methods in applied linguistics (pp. 302-311). Routledge.

[5] Bell, P., & Payant, C. (2020). Designing learner corpora: Collection, transcription, and annotation. In The Routledge handbook of second language acquisition and corpora (pp. 53-67). Routledge.

Co-constructing accessible meaning for the wider public: The role of interviewers in science popularization on health matters

Valeria Franceschi, **Sara Corrizzato**

University of Verona

Due to content-sharing and social media platforms, people are exposed to all kinds of information from a plethora of different sources, some of which are questionable in accuracy if not downright misleading or false (e.g. Kueffer and Larson 2014; Banks and DiMartino 2019). Hence, it is of paramount importance for academics and professionals to facilitate the dissemination of their research. Making scientific content understandable and clear to a non-expert public is a complex matter and scholars have explored the discourse of science popularization extensively in order to identify effective recontextualization (Calsamiglia and Van Djik, 2004). To date, however, science popularization in spoken dialogic form and through corpus linguistics methodology has been given little attention.

This study aims at investigating how interviewers in spoken science popularization communicative events contribute to increasing the comprehensibility and accessibility of scientific content by asking carefully targeted questions and clarifications to the speakers. To this end, the 'interviews' section of the SciencePop corpus (Facchinetti et al. 2024) was selected as a source of data. The subsection contains 26 broadcast interviews and 26 audio/video interviews on topics related to health and general well-being, for a total of around 218,868 running words. The analysis involves two steps: first, textual analysis software SketchEngine (Kilgarriff et al. 2014) is used to extract previously annotated questions posed by the interviewers, and in a second stage, the resulting concordance lines are closely analyzed to identify where disruption in communication may occur and how they are prevented or solved by interviewers, and through which strategies. Results of this research may contribute to the description of understudied popularizing genres as well as to the identification of comprehensibility issues in popularization events.

List of references

Banks D, Di Martino E (2019) "Introduction: Linguistic and Discourse Issues in Contemporary Scientific Communication. Aspects of Communicating Science to a Variety of Audiences". Journal of Pragmatics 139: 185e189.

Calsamiglia, H. and T. van Dijk. 2004. "Popularization Discourse and Knowledge about the Genome." Discourse & Society 15(4): 369–389.

Kueffer C and Larson MH (2014) "Responsible Use of Language in Scientific Writing and Science Communication". BioScience 64(8): 719-724.

Facchinetti R., Corrizzato S., Franceschi V. and Mambelli G. (2024). The SciencePop Corpus. Department of Foreign languages and Literatures, University of Verona.

Kilgarriff, A., Baisa V., Bušta J., Jakubíček, Kovář V., Michelfeit J., Rychlý P. and V.Suchomel. 2014. "The Sketch Engine: Ten Years on." Lexicography 1: 7-36.

The pragmatics of fraud: Trust management in the Enron Trader Tapes

Matteo Fuoli

University of Birmingham

Trust is essential for a thriving society, yet paradoxically, it also facilitates various forms of crime, from corporate fraud and bribery to romance scams and online grooming. While trust is built largely through verbal communication, the linguistic mechanisms underlying this process remain underexplored, particularly in criminal contexts. This study addresses this gap by proposing a new framework for studying how trust is built and manipulated in discourse and applying it to one of the most notorious cases of corporate misconduct in history: the Enron fraud.

We draw on forensic linguistics, applied psychology, and the 'move analysis' framework (Biber et al. 2007; Swales 1990; Upton & Cohen 2009) to analyze the Enron Trader Tapes Corpus (ETTC), a set of 505 phone conversation transcripts involving Enron employees during the 2000-2002 California energy crisis released as part of the legal proceedings against the company. The analysis uncovers the discursive mechanisms through which Enron managed trust both internally and externally while manipulating California's energy markets. The findings not only provide novel insights into the Enron case but also advance our understanding of the linguistic and pragmatic foundations of trust and the relationship between discourse, trust, and corporate corruption.

The ETTC, which we have compiled for this study and made publicly accessible for research purposes, represents a major new resource for forensic linguistics and discourse analysis. Approximately 415,000 words in size, it is the largest known spoken language corpus derived from a context of known illegal activity. Access to such clandestine conversations is rare, and existing datasets are often limited in scope, making the ETTC an invaluable asset for both this specific case study and forensic linguistics more broadly.

List of references

Biber, D., Ulla, C., & Upton, T. A. (2007). Discourse on the move: Using corpus analysis to describe discourse structure. John Benjamins.

Swales, J. M. (1990). Genre analysis: English in academic and research settings. Cambridge University Press.

Upton, T. A., & Cohen, M. A. (2009). An approach to corpus-based discourse analysis: The move analysis as example. Discourse Studies, 11(5), 585–605.

The use of vague language across proficiency levels in L2 speaking: Implications for language assessment

Dana Gablasova, Luke Harding

Lancaster University

Higher proficiency in L2 speaking has often been associated with an increase in lexical resources and the ability to use more sophisticated (e.g. less frequent, more specialised) vocabulary (Qian & Lin, 2019). This assumption has been also reflected in rating scales of major English language tests (e.g. IELTS, TOEFL iBT and PTE Academic) which mention an increase in precise vocabulary and a decrease in vague language in scoring criteria used to distinguish higher proficiency levels. However, the ability to use vague language – "forms that are intentionally fuzzy, general, and imprecise" (Cutting, 2012) – is a crucial aspect of communicative competence, allowing speakers to express (complex) ideas fluently and to deal with uncertainty or gaps in their factual knowledge, especially in unplanned interaction (McCarthy, 2020). Vague language has been shown to be a common feature of L1 English in informal and formal contexts (Cutting, 2007). However, the use of vague language in L2 speaking, and its patterning across proficiency levels, remains under-researched.

In this study, we used data from the Aptis Corpus which contains 1M words from the spoken component of the Aptis General exam (O'Sullivan, 2020) representing L2 production from 1,448 L2 speakers at four levels of proficiency - CEFR A2, B1, B2 and C. Data were searched for linguistic markers of vague language including 'thing', 'stuff', 'and so on' and 'kind of'. Frequency and type of vague language markers were analysed across proficiency levels. Results showed that the use of vague language markers increased systematically with proficiency. Textual analysis revealed that at higher levels of proficiency, the use of vague markers allowed speakers to – paradoxically - attain greater precision in their production and to express a wider range of meaning. The findings have important implications for challenging how lexical sophistication is understood in L2 proficiency scales.

List of references

Cutting, J. (Ed). (2007).Vague language explored. London: Palgrave Macmillan UK.

Cutting, J. (2012). Vague language in conference abstracts. Journal of English for academic purposes, 11(4), 283-293.

McCarthy, M. (2020). Vague language in business and academic contexts. Language Teaching, 53(2), 203-214.

O'Sullivan, B., Dunlea, J., Spiby, R., Westbrook, C., & Dunn, K. (2020). Aptis General technical manual, v 2.2. London: British Council.

Qian, D. D., & Lin, L. H. (2019). The relationship between vocabulary knowledge and language proficiency. The Routledge handbook of vocabulary studies (pp. 66-80).

Understanding Foundation-level ESL Students' academic writing: A corpus-based analysis of lexical collocation use

Mingyan Gao

University of Exeter

A key challenge for EFL/ESL learners in the process of learning and using native-like language is the written production of collocations. While extensive research (e.g., Li & Schmitt, 2010; Paquot, 2010; Sawaguchi & Mizumoto, 2022) has examined L2 learners' collocation use in academic writing across various contexts, the context of Foundation-level students remains under-researched. In English Medium Instruction contexts, Foundation Programmes aim to develop L2 students' academic literacy skills as pathways to undergraduate study. Understanding collocation use in this transitional stage can inform EAP teaching and support novice learners. This study investigates the use of lexical items from the Academic Collocation List (ACL; Ackermann & Chen, 2013) in Foundation student writing, comparing it to undergraduate student writing.

This paper will discuss two research questions: (1) how does ACL collocation use in Foundation writing compare to undergraduate writing? (2) to what extent are these collocations discipline-specific in Foundation writing? We analysed a learner corpus comprising 518 coursework assignments (c. 28 million words) written by 268 students in two cohorts of the Foundation Programme at a British institution. A reference corpus of 2,164 undergraduate texts selected from BAWE (levels 1-3) was used for comparison. Frequency, coverage, and dispersion analyses of 2,464 ACL entries were conducted using R packages for text cleaning and collocation extraction.

Results showed that ACL items had high coverage in Foundation writing (1.6%), with certain collocations frequently repeated, suggesting that Foundation students might over-rely on a narrower range of collocations, potentially reflecting a restricted lexical repertoire at this stage of their academic development. Analysis also revealed that lexical collocation use in Foundation writing was skewed towards Social Sciences. These insights underscore the need for targeted, discipline-specific instruction in Foundation Programmes to address lexical gaps and better prepare students for academic success.

List of references

Ackermann, K., & Chen, Y.-H. (2013). Developing the Academic Collocation List (ACL)–A corpus-driven and expert-judged approach. Journal of English for Academic Purposes, 12(4), 235–247.

Li, J., & Schmitt, N. (2010). The development of collocation use in academic texts by advanced L2 learners: A multiple case study approach. Perspectives on Formulaic Language: Acquisition and Communication, 22–46.

Paquot, M. (2010). Academic vocabulary in learner writing: From extraction to analysis. Continuum.

Sawaguchi, R., & Mizumoto, A. (2022). Exploring the use of make + noun collocations by Japanese EFL learners through a bilingual essay corpus. Corpora, 17(Supplement), 61–77. https://doi.org/10.3366/cor.2022.0247

Prompting corpus analysis through dynamic topic modeling: Preliminary results on a popular music album reviews corpus

Gilberto Giannacchi[1], **Sergio Picascia**[2]

[1]University of Insubria; [2]University of Milan

This paper shows the preliminary results obtained with the dynamic topic model BERTopic (Grootendorst 2022) on a popular music album reviews corpus, in a diachronic perspective (1980-2022). Despite the cultural relevance of the music press (Jones 2002, Grafe and McKeown 2024), its signature texts have been underrepresented in text and corpus linguistics. In this professional environment, album reviews enact a specific kind of meta-discourse, with critics reflecting on the discursive values of popular music (Van Leeuwen 2012). This discourse about music is based on a set of shared values between reviewers and audiences (Frith 1983, Shuker 2001). Dynamic topic modeling can provide "starting points for exploring […] underlying co-occurrence patterns or their implications" (Murakami et al. 2024: 1-2), as well as their diachronic evolution. BERTopic was used on a specialized, diachronic corpus consisting of album reviews published between 1980 and 2022 in the US and UK (2,681 texts; 1,691,293 tokens), built following the methods in Egbert, Biber, and Gray (2022). BERTopic's clustering techniques and its class-based variation of TF-IDF can allow for a coherent representation of topics (Grootendorst 2022). The model was tuned according to best practices reported in the documentation of the software (n.a., n.d., https://shorturl.at/lOdgi). BERTopic generated 527 topics, each one consisting of 10 semantically related n-grams – tri-grams as the maximum length. After a preliminary investigation, topics in the corpus were divided into four broad categories: musical instruments and production, musical genres, social actors, and outsiders – i.e., topics that have no explicit semantic connection to music (e.g., Topic 80: allegory, revelations, xenophobia, death, heaven, endings, epiphanies, existence, heavenly, life). We argue that these topics do not constitute musical meta-discourses in themselves. Rather, they can lead to the discovery of ever-shifting discursive patterns, which still require human intuition and corpus linguistics tools to be properly contextualized.

List of references

Egbert, J., Biber, D., and Gray, S. (2022). Designing and Evaluating Language Corpora: A Practical Framework for Corpus Representativeness. Cambridge: Cambridge University Press.

Frith S. (1983). Sound Effects: Youth, Leisure and the Politics of Rock'n'Roll. London: Constable.

Grafe, A. and McKeown, A. (eds.) (2024). Ink On The Tracks: Rock and Roll Writing. London: Bloomsbury.

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794.

Jones, S. (ed.) (2002). Pop Music and The Press. Philadelphia: Temple University Press.

Leeuwen, T.V. (2012). Critical Analysis of Multimodal Discourse. In The Encyclopedia of Applied Linguistics, C.A. Chapelle (Ed.). https://doi.org/10.1002/9781405198431.wbeal0269

Murakami, A., Thompson, P., Hunston, S., & Vajn, D. (2017). 'What is this corpus about?': using topic modelling to explore a specialised corpus. Corpora, 12(2), 243-277.

Shuker, R. (2001). Understanding Popular Music (2nd ed.). London and New York: Routledge.

What does human-like conversation look like in instant messaging? – A comparison between real and crowdworker-generated conversations –

Yelin Go[1], Jun Choi[2]

[1]CHONNAM NATIONAL UNIVERSITY; [2]CHONNAM NATIONAL UNIVERSITY

Since “Attention Is All You Need” (Vaswani et al., 2017), the introduction of “attention” mechanisms has revolutionized the paradigm of language understanding and generation in artificial intelligence, enabling machines to speak in a truly human-like manner. However, conversations between humans and machines still fall short of being equivalent to conversations between humans. Human conversations, such as those with ChatGPT, do not consist of neatly alternating, fully-formed utterances. Yet we still lack concrete knowledge of what constitutes a typical messenger conversation.

Messenger-based conversations are not only a relatively under-discussed register, but also come with challenges for large-scale data collection due to their private nature. As a result, recent training corpora for conversational AI increasingly rely on conversations generated by crowdworkers. However, no standardized method exists for evaluating how close these generated conversations are to actual human conversations.

In this context, this study analyzes real Korean messenger conversations between close acquaintances and compares them with crowdworker-generated messenger conversations (each with approximately 300,000 words). The comparison focuses on both morpho-syntactic and discourse-pragmatic aspects. On the morpho-syntactic level, we examine usage patterns of function words. On the discourse-pragmatic level, we compare features such as the average utterance length per turn and the use of discourse markers at turn transitions.

This study contributes to a better understanding of messenger conversations and supports the development and augmentation of training data for more natural conversations. Furthermore, this study points to the need for future research to move beyond making machines speak like humans and begin addressing what it means for them to converse like humans.

List of references

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

A comparison of lexical features in children's video media and child-directed speech

Anna Gowenlock, Jennifer Rodd, Beth Malory, Courtenay Norbury

University College London

Children's vocabulary knowledge is linked to the quantity and quality of language input they experience. By studying the lexical properties of different language sources, we can understand more about how children develop important vocabulary skills that support their success at school and beyond. Recent corpus studies point to children's books as an important source of diverse and sophisticated vocabulary that children are unlikely to encounter through day-to-day conversations with caregivers[1,2]. This project extends this work to another important type of language input: video media. Our goal was to understand how the lexical properties of children's video media differ from child-directed speech (CDS). We created a corpus of transcripts of programmes that are popular among 3-5-year-olds (~230,000 words) and compared this to CDS data from the CHILDES database (~2,590,000 words). In each corpus we examined features of lexical richness that are important to vocabulary acquisition. We found that the vocabulary in the video corpus is more diverse, sophisticated, and richer in meaning than the vocabulary in the CDS corpus. We also identified a set of keywords in each corpus and compared them on a number of psycholinguistic variables. Video keywords had a higher age of acquisition than CDS keywords, but there were no meaningful differences in concreteness or emotional variables between the keyword sets. Taken together, our findings suggest that video media could be a helpful source of lexical input for young children. Our results mirror findings about book language, suggesting that rich and diverse vocabulary may arise as a property of storytelling rather than being linked to a specific medium.

List of references

1. Dawson, N., Hsiao, Y., Tan, A., Banerji, N., & Nation, K. (2021). Features of lexical richness in children's books: Comparisons with child-directed speech. Language Development Research. https://doi.org/10.34842/5WE1-YK94
2. Montag, J. L., Jones, M. N., & Smith, L. B. (2015). The Words Children Hear: Picture Books and the Statistics for Language Learning. Psychological Science, 26(9), 1489–1496. https://doi.org/10.1177/0956797615594361

How do translators deal with repeated reporting verbs in literary texts? A multifactorial comparative study in English-to-Russian and English-to-Slovak language pair

Lukasz Grabowski[1], Daniel Borysowski[2], Filip Kalas[3], Lorenzo Mastropierro[4]

[1]University of Opole; [2]University of Opole; [3]University of Economics in Bratislava; [4]University of Insubria

In this multifactorial study, we aim to identify the predictors of repetition or lexical variety in the translation of reporting verbs from English into Russian and from English into Slovak. Using a sample of 20 literary novels (English-to-Russian) and 14 literary novels (English-to-Slovak) from InterCorp v. 15 (Rosen et al. 2022), we fit multiple negative binomial regression with mixed effects to assess the effect that selected predictor variables (e.g. frequency of a ST verb, its number of senses in Princeton WordNet) have on the response variable: the number of TT reporting verb types (lemmas) a ST reporting verb is translated into. If the number of types is high then it means that translators opted for lexical variety (i.e. used various TT reporting verbs as translation equivalents).

Our initial findings show that the overall model fit per the lowest AIC and BIC values obtained through backward elimination reveals that semantic category of a ST reporting verb, its frequency and translation date as well as the translator as a random effect have the largest individual contributions to explaining the proportion of variation (75%) in the response variable in the Russian translations, that is, the number of different verb types a ST verb is translated into. The low variance (0.05) in the random effect means that the impact of individual translators is relatively similar: there is some variability between the translators, but it is relatively small, and no single translator significantly influenced overall results. We later compare these findings with the ones in the English-Slovak language pair, where the model allowed us to explain 70% of variation in the response variable. Thus, the findings offer an attempt at explanation for the translator's choices in rendering recurring reporting verbs signalling direct speech, which carry important stylistic weight in literary texts.

List of references

Rosen, Alexandr, Martin Vavřín and Jan A. Zasina. 2022. The InterCorp Corpus – Czech), version 15 of 11 November 2022. Institute of the Czech National Corpus, Charles University, Prague 2022.

Evolution of registers as cultural constructs: The case of blogs

Marianna Gracheva[1], **Daniel Keller**[2], **Jesse Egbert**[3]

[1]Friedrich Alexander University Erlangen-Nürnberg; [2]Western Kentucky University; [3]Northern Arizona University

In text-linguistics, registers are culturally recognized text varieties, associated with the situation of use. Register research has now documented the existence of functional links between the situational characteristics of individual texts within registers and linguistic variation among those texts (Biber & Egbert, 2023; Egbert et al., 2024). These findings raise new questions about the evolution of registers as cultural constructs. First, as situations evolve, language must reflect language users' adaptations to new situations. Second, the degree of variation among texts at different points of a register's existence could reflect language users' degrees of convergence (or lack thereof) on communicative and linguistic register norms.

We examine these possibilities in blogs—a register characterized by a rapidly evolving technological landscape (Miller & Shepherd, 2009), whose life cycle—from the late 90s to the present day—is available for study. We analyze a new corpus of blogs from Blogspot.com spanning the years 1999–2023 (N_{texts}=2,452; N_{tokens}=~4,000,000; approx.100 texts/year). Our research questions are:

1. What linguistic features of blogs have become more or less frequent over time?
2. Have the linguistic features of blogs become more or less stable over time?

To address RQ1, we compute rates of occurrence for 150 lexico-grammatical features (Biber, 1988) and employ corresponding feature analysis (Egbert, 2024) to examine correlations between language use and time. The analysis reveals that features of oral/interactive communication and present orientation have increased over time, while features of information density, abstractedness, and past time references have decreased.

To address RQ2, we show change in linguistic stability—degree of linguistic variation—using several approaches, including coefficients of variation, borrowed from psycholinguistics (Segalowitz & Segalowitz, 1993), and relative standard error (e.g., Egbert et al., 2022, p.137). We conclude by discussing the implications of this research for blogs, specifically, and for diachronic and synchronic register studies, generally.

List of references

References:
Biber, D. (1988). Variation across speech and writing. Cambridge University Press.
Biber, D., & Egbert, J. (2023). What is register? Register Studies, 5(1), 1–22.
Egbert, J. (2024, September 14). The text-linguistic (r)evolution [Plenary talk]. American Association for Corpus Linguistics. Eugene, OR, USA.
Egbert, J., Biber, D., & Gray, B. (2022). Designing and evaluating language corpora: A practical framework for corpus representativeness. Cambridge University Press.
Egbert, J., Biber, D., Keller, D., & Gracheva, M. (2024). Register and the dual nature of functional correspondence: accounting for text-linguistic variation between registers, within registers, and without registers. Corpus Linguistics and Linguistic Theory.
Miller, C. R., & Shepherd, D. (2009). Questions for genre theory from the blogosphere. In J. Giltrow & D. Stein (Eds.), Genres in the internet: Issues in the theory of genre (pp. 263–290). John Benjamins.
Segalowitz, N. S., & Segalowitz, S. J. (1993). Skilled performance, practice, and the differentiation of speed-up from automatization effects: Evidence from second language word recognition. Applied Psycholinguistics, 14(3), 369–385.

Statistical learning in L2 non-adjacent verb-argument constructions

Jiaqi Feng Guo[1], **Pascual Pérez-Paredes**[2]

[1]University of Turku; [2]University of Murcia

Statistical learning plays a crucial role in second language acquisition, yet its operation in non-adjacent constructions remains underexplored. This study examines how learners acquire Mandarin verb-argument constructions with the preposition *duì* through a corpus-based investigation of frequency and contingency effects. We analyze whether learners' verb usage within these non-adjacent constructions reflects the distributional patterns found in their natural language input. Our analysis reveals that learners' usage patterns align with target language distributional regularities, providing strong evidence for statistical learning mechanisms in complex syntactic acquisition. Beyond distributional factors, we identify critical variables shaping input exposure, including accessibility, proficiency, and prototypicality, that illuminate L2 production choices. Through mixed-effects negative binomial regression, we demonstrate how advanced statistical modelling can effectively capture the complex variability inherent in linguistic data. This research not only advances our understanding of statistical learning in non-adjacent constructions but also highlights the importance of comprehensive approaches to L2 acquisition research. Our findings have important implications for both theoretical frameworks in statistical learning and practical applications in language pedagogy.

List of references

Alzahrani, A. (2021). The Effects of Two Association Measures on L2 Collocation Processing. International Journal of English Linguistics, 11(5), 28. https://doi.org/10.5539/ijel.v11n5p28

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language, 59(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Bley-Vroman, R. (2002). Frequency in production, comprehension, and acquisition. Studies in Second Language Acquisition, 24(2), 209–213.

Boone, G., Wilde, V. D., & Eyckmans, J. (2022). A longitudinal study into learners' productive collocation knowledge in L2 German and factors affecting the learning. Studies in Second Language Acquisition. https://doi.org/10.1017/s0272263122000377

Brooks, P. J., Kwoka, N., & Kempe, V. (2017). Distributional Effects and Individual Differences in L2 Morphology Learning. Language Learning, 67(1), 171–207. https://doi.org/10.1111/lang.12204

Bucholtz, M., & Hall, K. (2005). Identity and interaction: A sociocultural linguistic approach. Discourse Studies, 7(4–5), 585–614.

Bybee, J. L. (2010). Language, Usage and Cognition. CUP.

Cadierno, T., & Eskildsen, S. W. (2015). Usage-Based Perspectives on Second Language Learning. De Gruyter.

Candarli, D. (2021). A longitudinal study of multi-word constructions in L2 academic writing. Reading and Writing, 34(5), 1191–1223. https://doi.org/10.1007/s11145-020-10108-3

Che, W., Feng, Y., Qin, L., & Liu, T. (2021). N-LTP: An Open-source Neural Chinese Language Technology Platform with Pretrained Models. arXiv:2009.11616 [Cs]. http://arxiv.org/abs/2009.11616

Chen, C. (2002). 介词与介引功能 [Prepositions and its Introduction Function]. Anhui Education Press.

Chen, H., & Xu, H. (2019). Quantitative linguistics approach to interlanguage development. Lingua, 230, 102736. https://doi.org/10.1016/j.lingua.2019.102736

Clahsen, H., & Felser, C. (2006). Grammatical processing in language learners. Applied Psycholinguistics, 27(1), 3–42. https://doi.org/10.1017/S0142716406060024

Croft, W., & Cruse, D. A. (2004). Cognitive linguistics. CUP.

Crossley, S. A., Skalicky, S., Kyle, K., & Monteiro, K. (2019). Absolute frequency effects in second language lexical acquisition. Studies in Second Language Acquisition, 41(4), 721–744.

DeKeyser, R. (2007). Practice in a Second Language: Perspectives from Applied Linguistics and Cognitive Psychology. CUP. https://doi.org/10.1017/CBO9780511667275

Deshors, S. C., & Gries, S. T. (2022). Using Corpora in Research on Second Language Psycholinguistics. In The Routledge Handbook of Second Language Acquisition and Psycholinguistics. Routledge.

Eckerth, J., & Tavakoli, P. (2012). The effects of word exposure frequency and elaboration of word processing on incidental L2 vocabulary acquisition through reading. Language Teaching Research, 16, 227–252. https://doi.org/10.1177/1362168811431377
Edmonds, A., & Gudmestad, A. (2023). Phraseological Use and Development During a Stay Abroad. Language Learning, 73(2), 475–507. https://doi.org/10.1111/lang.12547
Elder, C., McNamara, T., Kim, H., Pill, J., & Sato, T. (2017). Interrogating the construct of communicative competence in language assessment contexts. Language & Communication, 57, 14–21. https://doi.org/10.1016/J.LANGCOM.2016.12.005
Ellis, N. C., & Ferreira–Junior, F. (2009). Construction Learning as a Function of Frequency, Frequency Distribution, and Function. Modern Language Journal, 93(3), 370–385. https://doi.org/10.1111/j.1540-4781.2009.00896.x
Ellis, N. C., Römer, U., O'Donnell, M. B., & Schleppegrell, M. J. (2016). Usage-Based Approaches to Language Acquisition and Processing. Wiley.
Erickson, L. C., & Thiessen, E. D. (2015). Statistical learning of language: Theory, validity, and predictions. Developmental Review, 37, 66–108. https://doi.org/10.1016/j.dr.2015.05.002
Eskildsen, S. W. (2022). Usage-Based SLA: From Corpora to Social Interaction. In The Routledge Handbook of Second Language Acquisition and Sociolinguistics. Routledge.
Gablasova, D., Brezina, V., Mcenery, T., & Boyd, E. (2017). Epistemic Stance in Spoken L2 English. Applied Linguistics, 38(5), 613–637. https://doi.org/10.1093/applin/amv055
Gass, S. M., Behney, J., & Plonsky, L. (2020). Second language acquisition: An introductory course. Routledge.
Gass, S. M., & Mackey, A. (2002). Frequency effects and second language acquisition. Studies in Second Language Acquisition, 24(2), 249–260.
Gelman, A., Hill, J., & Yajima, M. (2012). Why We (Usually) Don't Have to Worry About Multiple Comparisons. Journal of Research on Educational Effectiveness, 5(2), 189–211. https://doi.org/10.1080/19345747.2011.618213
Goldberg, A. E. (1995). Constructions: A construction grammar approach to argument structure. University Chicago Press.
Goldberg, A. E., Casenhiser, D. M., & Sethuraman, N. (2004). Learning argument structure generalizations. Cognitive Linguistics, 15(3), 289–316. https://doi.org/10.1515/cogl.2004.011
Granena, G., & Long, M. H. (2013). Age of onset, length of residence, language aptitude, and ultimate L2 attainment. Second Language Research, 29(3), 311–343. https://doi.org/10.1177/0267658312461497
Gries, S. T. (2010). Useful statistics for corpus linguistics. A Mosaic of Corpus Linguistics: Selected Approaches, 66, 269–291.
Gries, S. Th. (2022). Toward more careful corpus statistics. Research Methods in Applied Linguistics, 1(1), 100002. https://doi.org/10.1016/j.rmal.2021.100002
Hashimoto, B. J., & Egbert, J. (2019). More Than Frequency? Language Learning, 69(4), 839–872. https://doi.org/10.1111/lang.12353
Hilpert, M. (2014). Collostructional analysis. Corpus Methods for Semantics: Quantitative Studies in Polysemy and Synonymy, 43, 391.
Hoey, M. (2012). Lexical priming: A new theory of words and language. Routledge.
Hopp, H. (2023). Sentence processing in a second language: Linguistic approaches. The Routledge Handbook of Second Language Acquisition and Psycholinguistics, 216–228.
Hulstijn, J. H., Ellis, R., & Eskildsen, S. (2015). Orders and Sequences in the Acquisition of L2 Morphosyntax. Language Learning, 65. https://doi.org/10.1111/lang.12097
Ibbotson, P. (2013). The Scope of Usage-Based Theory. Frontiers in Psychology, 4. https://doi.org/10.3389/fpsyg.2013.00255
Johnson, R. A. (2009). Statistics: Principles and Methods (6th ed.). Wiley.
Kartal, G., & Sarigul, E. (2017). Frequency Effects in Second Language Acquisition. Journal of Education and Training Studies, 5(6), 1. https://doi.org/10.11114/jets.v5i6.2327
Kyle, K. (2021). Natural language processing for learner corpus research. International Journal of Learner Corpus Research, 7(1), 1–16. https://doi.org/10.1075/ijlcr.00019.int
Lenth, R. (2022). emmeans: Estimated Marginal Means, aka Least-Squares Means (R package version 1.8.0). https://CRAN.R-project.org/package=emmeans
Li, L. (1999). 介词"对"的意义和用法考察 [An Investigation of the Meaning and Usage of Preposition 'dui']. Journal of Tianjin Normal University, 4, 71–75.
Marinis, T., Roberts, L., Felser, C., & Clahsen, H. (2005). Gaps in second language sentence processing. Studies in Second Language Acquisition, 27(1), 53–78.

Martinez, R., & Murphy, V. A. (2011). Effect of Frequency and Idiomaticity on Second Language Reading Comprehension. TESOL Quarterly, 45(2), 267–290. https://doi.org/10.5054/tq.2011.247708
McElreath, R. (2020). Statistical Rethinking: A Bayesian Course with Examples in R and STAN (2nd ed.). CRC Press.
McEnery, T., & Brezina, V. (2022). Fundamental principles of corpus linguistics. CUP.
Meurers, D. (2015). Learner corpora and natural language processing. The Cambridge Handbook of Learner Corpus Research, 537–566.
Monteiro, K. R., Crossley, S. A., & Kyle, K. (2020). In search of new benchmarks. Applied Linguistics, 41(2), 280–300.
Murakami, A., & Ellis, N. C. (2022). Effects of Availability, Contingency, and Formulaicity on the Accuracy of English Grammatical Morphemes in Second Language Writing. Language Learning. https://doi.org/10.1111/lang.12500
Nation, I. S. P. (2013). Learning Vocabulary in Another Language (2nd ed.). CUP. https://doi.org/10.1017/CBO9781139858656
Paquot, M., & Gries, S. T. (2020). A Practical Handbook of Corpus Linguistics. Springer.
Pellicer-Sánchez, A. (2016). Incidental L2 Vocabulary Acquisition From and While Reading. Studies in Second Language Acquisition, 38(1), 97–130. https://doi.org/10.1017/S0272263115000224
Pérez-Paredes, P. (2020). Corpus Linguistics for Education: A Guide for Research. Routledge.
Peters, E. (2018). The effect of out-of-class exposure to English language media on learners' vocabulary knowledge. ITL - International Journal of Applied Linguistics, 169(1), 142–168. https://doi.org/10.1075/itl.00010.pet
Rebuschat, P., & Williams, J. N. (2012). Statistical Learning and Language Acquisition. De Gruyter.
Römer, U., & Garner, J. (2019). The development of verb constructions in spoken learner English. International Journal of Learner Corpus Research, 5(2), 207–230. https://doi.org/10.1075/ijlcr.17015.rom
Sebastián-Gallés, N., Echeverría, S., & Bosch, L. (2005). The influence of initial exposure on lexical representation. Journal of Memory and Language, 52(2), 240–255.
Siyanova, A., & Schmitt, N. (2008). L2 Learner Production and Processing of Collocation. Canadian Modern Language Review, 64(3), 429–458. https://doi.org/10.3138/cmlr.64.3.429
Sorace, A. (2006). Possible manifestations of shallow processing in advanced second language speakers. Applied Psycholinguistics, 27, 88–91. https://doi.org/10.1017/S0142716406060164
Stutterheim, C. V., Lambert, M., & Gerwien, J. (2021). Limitations on the role of frequency in L2 acquisition. Language and Cognition, 13(2), 291–321. https://doi.org/10.1017/langcog.2021.5
Su, X. (2013). 现代汉语语义分类词典 [A Thesaurus of Modern Chinese]. Commercial Press.
Tono, Y. (2004). Multiple comparisons of IL, L1 and TL corpora: The case of L2 acquisition of verb subcategorization patterns by Japanese learners of English. Corpora and Language Learners, 45.
Trahey, M., & White, L. (1993). Positive evidence and preemption in the second language classroom. Studies in Second Language Acquisition, 15(2), 181–204.
Uchihara, T., Webb, S., & Yanagisawa, A. (2019). The Effects of Repetition on Incidental Vocabulary Learning. Language Learning. https://doi.org/10.1111/LANG.12343
Winke, P., Gass, S., & Sydorenko, T. (2013). Factors Influencing the Use of Captions by Foreign Language Learners. Modern Language Journal, 97(1), 254–275. https://doi.org/10.1111/j.1540-4781.2013.01432.x
Winter, B., & Bürkner, P.C. (2021). Poisson regression for linguists: A tutorial introduction to modelling count data with brms. Language and Linguistics Compass, 15(11), e12439.
Wolter, B., & Gyllstad, H. (2013). Frequency of input and L2 collocational processing. Studies in Second Language Acquisition, 35(3), 451–482.
Wulff, S. (2020). Usage-based Approaches. In The Routledge Handbook of Second Language Acquisition and Corpora. Routledge.
Xun, E., Rao, G., Xiao, X., & Zang, J. (2016). The Construction of the BCC-Corpus in the Age of Big Data. Corpus Linguistics, 3(1), 93–109.
Yan, J., & Liu, H. (2022). Semantic Roles or Syntactic Functions: The Effects of Annotation Scheme on the Results of Dependency Measures. Studia Linguistica, 76(2), 406–428. https://doi.org/10.1111/stul.12177
Zhou, X. (2007). 介词的语法性质和介词研究的系统方法 [Grammatical Features of Chinese Prepositions and a Systematic Method in Preposition Study]. Zhongshan University Journal, 3, 109–115.

Death and desire: A corpus approach to vampire erotica

Kat Gupta[1], **Alon Lischinsky**[2]

[1]Royal Holloway, University of London; [2]Oxford Brookes University

The figure of the vampire has played a key role in modern popular culture as a powerful symbol of sexuality and, simultaneously, of the danger, degradation and disgust that can be attached to sexual attraction (Hobson, 2016, p. 11; Rowe, 1995, p. 163). And as sexuality — including non-normative sexuality — has been afforded more space in the public sphere, such figures have gained greater visibility and variety (Day, 2002, pp. 25, 78).

Gender and culture studies have produced a rich literature exploring the erotic potential of the vampire in romance fiction (e.g., Crawford, 2014, p. 14), film (e.g., Ní Fhlainn, 2019), TV (e.g., Anyiwo, 2016) or video games (e.g., Escandell Montiel & Borham Puyal, 2020). Much of this research notes how vampire media often skirt the pornographic, yet shies from texts that are actually explicit. Scholars have noted the increasing popularity of vampire porn both visual and written (Marks, 2014, 2018), but analyses of its content, structure and conventions are very rare (a ground-breaking exception can be found in Bosky, 1999).

In this paper, we explore the characterisation of vampires and the narrative roles they play in a corpus of approximately 6500 stories collected from Literotica.com (2016), one of the oldest, largest and most widely-read erotic fiction repositories online. We use keyword analysis to demonstrate how the distinctive vocabulary of these stories shows their narrative focus on the vampire's gaze, voice and body, but also how these are used to index issues of power, domination and gendered violence.

We argue that the desublimation of the sexual dimensions of the vampire myth has shifted it from an image of the dangerous and despised other to a cipher of a sexual self that is fundamentally dangerous to the social order, especially in the ways it troubles naive concepts of gender.

List of references

Anyiwo, U. M. (2016). Beautifully Broken. In A. Hobson & U. M. Anyiwo (Eds.), Gender in the Vampire Narrative (pp. 93–108). SensePublishers. https://doi.org/10.1007/978-94-6300-714-6_7

Bosky, B. L. (1999). Making the Implicit, Explicit: Vampire Erotica and Pornography. In L. G. Heldreth & M. Pharr (Eds.), The Blood is the Life: Vampires in Literature (pp. 217–234). Popular Press.

Crawford, J. (2014). The Twilight of the Gothic?: Vampire Fiction and the Rise of the Paranormal Romance, 1991-2012. University of Wales Press. https://books.google.com/books?hl=en&lr=&id=0mGuBwAAQBAJ&oi=fnd&pg=PP1&ots=zptlfVWcDe&sig=En8-OTezVUyyry7WVh4RTr1mR7U

Day, W. P. (2002). Vampire legends in contemporary American culture: What becomes a legend most. University Press of Kentucky.

Escandell Montiel, D., & Borham Puyal, M. (2020). Villains and Vixens: The Representation of Female Vampires in Videogames. Oceánide, 12, 85–93. https://doi.org/10.37668/oceanide.v12i.29

Hobson, A. (2016). Dark Seductress: The Hypersexualization of the Female Vampire. In A. Hobson & U. M. Anyiwo (Eds.), Gender in the Vampire Narrative (pp. 9–27). SensePublishers. https://doi.org/10.1007/978-94-6300-714-6_2

Marks, L. H. (2014). “I Eat Brains … or Dick” Sexual Subjectivity and the Hierarchy of the Undead in Hardcore Film. In S. McGlotten & S. Jones (Eds.), Zombies and sexuality: Essays on desire and the living dead (pp. 159–179). McFarland & Company.

Marks, L. H. (2018). Alice in pornoland: Hardcore encounters with the Victorian gothic. University of Illinois Press.

Ní Fhlainn, S. (2019). Postmodern Vampires: Film, Fiction, and Popular Culture. Palgrave Macmillan UK. https://doi.org/10.1057/978-1-137-58377-2

Rowe, M. (1995). Writing below the belt: Conversations with erotic authors. Masquerade Books.

Quandaries of corpus analysis in the age of AI hype

Andrew Hardie

Lancaster University

Since late 2022, so-called Generative Artificial Intelligence (AI) systems based on Large Language Models (LLMs) that apparently offer serious advances on flawed systems of yesteryear – e.g. ChatGPT versions 3 and 4 – have penetrated the zeitgeist of both academia and the culture at large. Pundits predict major impacts on business, economy, and society. Academics, and other educators, have been urged to modify teaching practices in light of the potential for AI-assisted cheating. Only the spectre of AI 'hallucinating' non-existent facts remains to deter such subterfuge.

For corpus linguists, especially those concerned with language pedagogy, tantalising prospects may seem to arise from this new AI wave. Can generative AI complement corpus methods by its similar ability to search and summarise massive text collections? By nature an LLM parameterises a dataset on the same kind of scale as a language corpus (millions or billions of tokens). And that machine-learning process can be seen as akin to well-known techniques of corpus analysis (concordances, collocation, keyness) that since c.1980 have allowed language teachers to build on empirically founded descriptions of the target language. We are being driven to ask, explicitly or implicitly: *can the new AI systems do the job I am currently doing with painstaking corpus analysis?*

I argue, from the perspective of a corpus methodologist, for scepticism. First, I will problematise current narratives regarding what AI is and how it works. Present systems are not truly 'artificial intelligence' as most people would understand the term. Second, I will address the specific prospect that the operation of LLMs can substitute in part or whole for a corpus linguist's analysis. This proposition is highly improbable due to the opacity and inscrutability of black-box algorithms in comparison to well-understood corpus methods, deployed with care and skill.

Exploring the genre of Land Acknowledgments: A corpus-based move analysis with ChatGPT integration

Jack A. Hardy[1], Savannah T. Brown[1], Ravi Parikh[2], Carl Yu[1], Sarah Bekele[3]

[1]Oxford College of Emory University; [2]Westminster Schools; [3]Independent Researcher

Land acknowledgments in U.S. higher education reflect efforts to recognize Indigenous histories (Author, 2024). This presentation describes how we examined the genre by combining the traditional genre analytical method of move analysis (Swales, 1990) with ChatGPT-assisted annotations. This study asks how land acknowledgment functional moves/steps are structured and how LLM tools can enhance analysis.

To that end, we compiled a corpus of 722 land acknowledgments (107,008 words) from U.S. colleges and universities and created a move/step framework, developed iteratively through human and ChatGPT collaboration. The automatic annotations were evaluated through a systematic comparison with those of human raters. In this presentation, we describe that iterative process as well as the final framework developed for analysis. The three moves (i.e., Establish historical context; Acknowledge land or legal significance; Demonstrate responsibility) and their respective steps are then described along with the linguistic markers that frequently align with specific rhetorical steps, which were used to help the GPT annotate the data. We then describe the coverage of each move and step in the data, highlighting which are relatively obligatory and which are outliers.

This talk highlights both methodological contributions and cultural insights. Methodologically, it builds on the work of Yu (2025), which demonstrates the feasibility of integrating ChatGPT into corpus-based genre analysis, offering efficiency and scalability with potential to incorporate critical nuance and human interpretation. Culturally, it contributes to the growing body of research on emerging institutional, performative genres and their socio-political implications.

Future directions include a more nuanced genre analysis (e.g., examining move/step orders; correlations between states and types of language used) as well as exploration of additional linguistic dimensions such as sentiment and speech acts. This work bridges computational and humanistic approaches, advancing our understanding of the evolving genre of land acknowledgments and the potential of ChatGPT to enhance linguistic inquiry.

List of references

Author (2024)

Swales, J. (1990). Genre analysis: English in academic and research settings. Cambridge University Press.

Yu, D. (2025). Towards LLM-assisted move annotation: Leveraging ChatGPT-4 to analyse the genre structure of CEO statements in corporate social responsibility reports. English for Specific Purposes, 78, 33-49. https://doi.org/10.1016/j.esp.2024.11.003

Using the NewsScape Corpus to explore multimodal meaning-making in TV news communication about immigration

Christopher Hart

Lancaster University

Corpus-assisted CDA has shown how refugees and migrants are constructed in online and print-news media, verbally, visually and multimodally (Gabrielatos & Baker 2008; Martínez Lirola 2017; Romano & Porto 2021). Owing to difficulties associated with obtaining and analysing large quantities of televisual data, however, the discursive construction of refugees and migrants in TV news has not been subject to similar interrogation. This paper exploits the NewsScape Corpus – a massive multimodal corpus of broadcast news collated by the Distributed Red Hen lab (https://www.redhenlab.org/) – to investigate the multimodal representation of refugees and migrants in television news. Accessed via CQPWeb (Hardie 2012), the corpus is searched for target utterances representing four different constructions: *refugees/*migrants have VERBed* and *refugees/*migrants are VERBing.* With a focus on expressions of motion and following filtering of the data to exclude noise, the co-verbal images accompanying 474 utterances are analysed quantitatively and qualitatively. Results show that refugees/migrants are depicted in large rather than small groups, that they are depicted in transit somewhere along the migratory journey rather than in countries of origin or destination countries, that they are depicted on land more than at sea, that they are depicted in security contexts and that they are erased represented instead through abstract forms such as maps and silhouettes. Certain of these language-image combinations emerge as obtaining genre-specific multimodal constructional status (Steen & Turner 2013) and whose visual component is thus likely to be evoked even when not co-instantiated in discourse. The ideological implications of these patterns of representation are discussed from the perspective of multimodal CDA (Machin 2013) where, for example, large-group depictions are shown to have dehumanising effects (Azavedo et al. 2021). The paper presents an empirical case study but also serves to demonstrate the utility of the NewsScape Corpus as a resource for multimodal corpus-assisted CDA.

List of references

Azevedo, R. T., De Beukelaer, S., Jones, I. L., Safra, L., and Tsakiris, M. (2021). When the lens is too wide: the political consequences of the visual dehumanization of refugees. Human. Soc. Sci. Commun. 8:115.

Gabrialetos, C., and Baker, P. (2008). Felling, sneaking, flooding: A corpus analysis of discursive constructions of refugees and asylum seekers in the UK press 1996-2005. J. English Linguist. 36, 5–38.

Hardie, A. (2012). CQPweb – combining power, flexibility and usability in a corpus analysis tool. Int. J. Corpus Linguist. 17, 380–409.

Machin, D. (2013). What is multimodal critical discourse studies? Crit. Disc. Stud. 10 (4), 347-355.

Martínez Lirola, M. (2016). Linguistic and visual strategies for portraying immigrants as people deprived of human rights. Soc. Semiot. 27, 21–38.

Romano, M., and Dolores Porto, M. (2021). “Framing CONFLICT in the Syrian refugee crisis: multimodal representations in the Spanish and British press” in Discursive approaches to sociopolitical polarization and conflict. eds. L. Filardo-Llamas, E. Morales-López, and A. Floyd (London: Routledge), 153–173.

Mapping grammatical variation in North American English using a YouTube corpus

Brett Hashimoto[1], **Joey Stanley**[1], **Jack Grieve**[2]

[1]Brigham Young University; [2]University of Birmingham

Research on variation in North American English has benefitted from increasingly larger data collections. However, studies on grammatical variation are either based on elicited data (Zanuttini et al., 2018) or written data that may have undergone (self-)editing (e.g., Grieve, 2016; Huang et al., 2016). We aim to uncover regional grammatical variation based on transcriptions of spoken data. The present study uses the 1.25-billion-word (301,846 texts from 2,572 YouTube channels, amounting to 154,041 hours of video) Corpus of North American Spoken English (CoNASE; Coats, 2019) to examine 45 grammatical features that have been documented to vary across North American dialects of English using a custom Python script (Grieve, 2016). Each text in the corpus was annotated for geospatial location. Each grammatical feature was counted to produce normalized rates of occurrence by county and a mean-variant-preference value was calculated, which represents the aggregated preference score of a county for a grammatical variant. Global Moran's I values were then calculated for each county to assess spatial autocorrelation. Then data underwent an Exploratory then Confirmatory Factor Analysis as a variable reduction method before it underwent a Hierarchical Cluster Analysis to group counties according to how the grammatical features are used. The results were then mapped using GIS software which allows for smoothed interpolation over the entire region. The results of this study produced maps of dialects according to grammatical features used in spoken data, which we then compared against previous sociolinguistic mapping research based on phonetic (Labov et al. 2006), lexical (Carver 1987), or written data (Grieve, 2016; Huang et al., 2016). The mapping of grammatical features largely aligns with other dialect maps of the United States; however, our preliminary analysis indicates several interesting differences, especially when using different numbers of clusters in the clustering solution, indicating levels of dialectal differences.

List of references

Carver, C. (1987). American regional dialect: A word geography. Ann Arbor: University of Michigan Press.

Coats, S. (2019). A corpus of regional American language from YouTube. In Proceedings of the Digital Humanities in the Nordic Countries 4th Conference Copenhagen, Denmark, March 5-8, 2019. RWTH Aachen University.

Grieve, J. (2016). Regional variation in written American English. Cambridge University Press.

Huang, Y., Guo, D., Kasakoff, A., & Grieve, J. (2016). Understanding US regional linguistic variation with Twitter data analysis. Computers, environment and urban systems, 59, 244-255.

Labov, W., Ash, S., & Boberg, C. (2006). The atlas of North American English: Phonetics, phonology and sound change. Mouton de Gruyter.

Zanuttini, Raffaella, et al. "The Yale Grammatical Diversity Project: Morphosyntactic variation in North American English." Linguistics Vanguard 4.1 (2018): 20160070.

Utilizing ChatGPT-o1 to compile a move corpus for three-minute thesis presentations: Developing a move-based phraseological list

Yohei Hirano[1], Tatsuya Ishii[2], Mark Andrew Pileggi[1]

[1]Kobe City College of Technology; [2]Kochi University

This study introduces a novel method for creating a move-based phraseological list by building a move corpus with the assistance of generative AI, specifically ChatGPT-o1. Based on the assumption that members of a discourse community share common logical structures and language behaviors, traditional corpus-based move analyses have relied on manual construction to identify frequent n-grams or phrase frames across various fields (e.g., Casal & Kessler, 2020; Liu & Chen, 2022; Lu & Kisselev, 2021). While integrating qualitative and quantitative methods is essential, the manual process remains labor-intensive. To address the challenge of defining move boundaries, pioneering studies have utilized ChatGPT-4 to analyze abstracts in research papers (Kim & Lu, 2024), introduction sections (Yu et al., 2024), and CEO statements in corporate social responsibility reports (Yu, 2025). In this research, we employed ChatGPT-o1to develop a move corpus for Three-Minute Thesis (3MT) presentations, a popular academic speaking format among doctoral students. We transcribed 200 presentations (totaling approximately 88,000 tokens) from 12 universities available on YouTube. ChatGPT-o1 was used to categorize the transcripts into the eight moves identified by Hu and Lui (2018). To prevent the overlap of highly frequent n-grams across the entire corpus (e.g., “a lot of”), we utilized CasualConc (Imao, 2024) to extract dispersion-based statistically significant three-grams to five-grams specific to each move corpus compared to the whole corpus, thereby extending Gries’s (2024) concept of key words. This approach effectively identified phraseological units strongly associated with each move, such as "we need to" in the Orientation move and "where my research comes in" in the Framework move. Although the inter-reliability of the move boundaries needs to be evaluated (Kim et al., 2024; Rau & Shih, 2021), this corpus-based move analysis using ChatGPT-o1 could pave the way for creating a domain-specific move-based phraseological list.

List of references

Gries. S. (2024). Frequency, Dispersion, Association, and Keyness: Revising and Tupleizing Corpus Linguistic. John Benjamins.

Kim, M., Qiu, X., & Wang, Y. (2024). Interrater agreement in genre analysis: A methodological review and a comparison of three measures. Research Methods in Applied Linguistics, 3(1), 1-17. https://doi.org/10.1016/j.rmal.2024.100097

Kim, M., & Lu, X. (2024). Exploring the potential of using ChatGPT for rhetorical move-step analysis: The impact of prompt refinement, few-shot learning, and fine-tuning. Journal of English for Academic Purposes, 71,1-15. https://doi.org/10.1016/j.jeap.2024.101422

Rau, G., & Shih, Y.-S. (2021). Evaluation of Cohen’s kappa and other measures of inter-rater agreement for genre analysis and other nominal data. Journal of English for Academic Purposes, 53, 1-11. https://doi.org/10.1016/j.jeap.2021.101026

Yu, D. (2025). Towards LLM-assisted move annotation: Leveraging ChatGPT-4 to analyse the genre structure of CEO statements in corporate social responsibility reports. English for Specific Purposes, 78, 33–49. https://doi.org/10.1016/j.esp.2024.11.003

Yu, D., Bondi, M., & Hyland, K. (2024). Can GPT-4 learn to analyse moves in research article abstracts? Applied Linguistics, 1-19. https://doi.org/10.1093/applin/amae071

"One does not necessarily need to understand satoori to know that it is hot": Attitudes toward stigmatised Korean regional dialects among English speakers on Twitter

Caitlin Hogan

Queen Mary University of London

This study uses corpus methods to explore dialect attitudes on social media—namely, perceptions of non-standard regional Korean varieties (*satoori*) amongst English speakers on Twitter. In South Korea, satoori is stigmatized in favour of Standard Korean, and while research on dialectology is not extensive, previous scholarship finds attitudes toward regional varieties in Korea characterise them as boorish, aggressive, or masculine (e.g., Jeon, 2013; Kang, 2015; Kwon & Bae, 2020). K-pop idols predominantly use Standard Korean in professional contexts. However, idols occasionally perform satoori, and these moments circulate globally via social media, including those with minimal understanding of Korean language ideologies.

The research questions are: (RQ1) What attitudes do English-speaking Twitter users hold towards satoori? (RQ2) How do these attitudes differ from domestic perceptions, as identified in prior research?

The study employs a quantitative corpus approach to analyse language attitudes (see Durham, 2016), leveraging a 1.2-million-word corpus (89,368 tweets) collected via the Twitter Academic API using the term satoori, spanning 2009–2022. Corpus-assisted discourse analysis (Partington et al., 2013) was applied to examine how satoori is framed among posters, combining analysis of the modification and predication of the term with qualitative analysis of discourses evoked in evaluations of the varieties.

Findings reveal that posters associated satoori with authenticity, emotional intensity, and attractiveness, and not the traits associated with satoori in Korea. It also emerges that fans use the term satoori to refer to other non-standard language varieties in different contexts (for example, to refer to dialects in Malaysia and Indonesia) as a way to perform their identity as a fan of Korean culture. This research suggests the utility of social media in the investigation of dialect attitudes. It also raises questions about the spread of Korean culture and the consequences for ideologies of the Korean language both abroad and domestically.

List of references

Durham, M. (2016). Changing Attitudes Towards the Welsh English Accent: A View from Twitter. In M. Durham & J. Morris (Eds.), Sociolinguistics in Wales (pp. 181–205). Palgrave Macmillan UK. https://doi.org/10.1057/978-1-137-52897-1_7

Jeon, L. (2013). Drawing Boundaries and Revealing Language Attitudes: Mapping Perceptions of Dialects in Korea. University of North Texas.

Kang, Y. J. (2015). Perceptions Of Korean Dialects by Gyeongsang Residents. San Diego State University.

Kwon, M., & Bae, Y. (2020). A dialect attitude study on Jeju dialect—Focusing on public officials in Jeju -. The Korean Association for Dialectology, 31, 183–217. https://doi.org/10.19069/kordialect.2020.31.183

Partington, A., Duguid, A., & Taylor, C. (2013). Patterns and Meanings in Discourse: Theory and practice in corpus-assisted discourse studies (CADS). John Benjamins Publishing.

Corpus-derived phraseological units and their effectiveness in estimating L2 writing quality

Ping-Yu Huang

Ming Chi University of Technology

Among corpus-derived phraseological units, *n*-grams have been the predominant measures adopted to assess L2 writing quality (e.g., Bestgen, 2017). Recently, Paquot (2018; 2019) introduced a more refined measure, dependency relation pairs (i.e., words connected by syntactic dependency relations such as verb + direct object), and empirically showed that higher-quality L2 texts did contain more strongly-associated dependency pairs. This finding was corroborated by Kyle & Eguchi (2021), who further demonstrated that dependency pairs, compared to bigrams, were stronger predictors of L2 writing scores. Building on this line of research, the current study examines the unique predictive value of dependency relations by comparing the predictive power of dependency pairs and non-dependency pairs (i.e., pairs without dependency relations) in relation to L2 writing quality. Furthermore, unlike the previous studies that tested only 3~4 dependency relations, we included and investigated a total of 17 relations. Two sets of learner texts were analyzed. The first consisted of 360 CEFR B2-level essays written by Taiwan learners of English, and the second comprised 640 CEFR A2~B2 essays from the ICNALE corpus (Ishikawa, 2023). All dependency and non-dependency pairs appearing within 2- or 3-word windows were extracted from the texts, and their strengths of association were determined using the COCA corpus (Davies, 2008–). The correlational and regression analyses performed, however, revealed inconsistent results. Both types of pairs effectively predicted the scores in the first dataset, but only the dependency ones accounted for the ICNALE scores. Further examination of the ICNALE data indicated that non-dependency pairs were predictive only at the B2 proficiency level. These results taken together suggest that different types of phraseological units would account for L2 writing quality at different proficiency levels, a finding not reported previously. We discuss our results based on frequency-based approaches to language learning, and offer implications for automated writing evaluation.

List of references

Bestgen, Y. (2017). Beyond single–word measures: L2 writing assessment, lexical richness and formulaic competence. System, 69, 65–78.

Davies, M. (2008–). The Corpus of Contemporary American English (COCA): 520 million words, 1990–present. Available online at http://corpus.byu.edu/coca/.

Ishikawa, S. (2023). The ICNALE Guide: An Introduction to a Learner Corpus Study on Asian Learners' L2 English (Routledge).

Kyle, K., & Eguchi, M. (2021). Automatically assessing lexical sophistication using word, bigram, and dependency indices. In S. Granger (Ed.), Perspectives on the L2 Phrasicon: The View from Learner Corpora. Multilingual Matters.

Paquot, M. (2018). Phraseological competence: A missing component in university entrance language tests? Insights from a study of EFL learners' use of statistical collocations. Language Assessment Quarterly, 15, 29–43.

Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. Second Language Research, 35, 121–145.

Estimation of relative frequency via Large Language Model

Weihang Huang, Jack Grieve, Akira Murakami

University of Birmingham

The relative frequency (RF) of a word type measures its rate of occurrence in one or more texts (McEnery et al., 2011). In corpus linguistics, RF is used (1) to describe the use of a type in one or more texts or (2) to predict the use of a type in a population of texts based on a sample of texts (i.e. corpus) drawn from that population.

In the second case, the calculation of RF is equivalent to maximum likelihood estimation (MLE). However, MLE is generally incorrect because types absent from the corpus will be assigned a RF of zero, although these types may occur in the population from which the corpus is drawn, while the RFs of types that do occur risk being overestimated.

To address this issue, estimators such as Add-One and Good Turing have been developed (Gale et al., 1994). For example, the Add-One approach adds one to the frequencies of all types in the sample before computing RFs. However, all these approaches assign types with the same frequency the same RF, even though we should not assume that all types that are absent from the sample have the same RFs in the population.

To address this issue, we propose LERF (LLM-based Estimated Relative Frequency), an estimator of RF based on negative log-likelihood values derived from large language models (LLMs). The advantage of LERF is that it uses LLMs pretrained on large general corpora to incorporate contextual information into RF estimation.

We test LERF using six LLMs (gpt2, gpt2-medium, gpt2-large, gpt2-xl, llama-2-7b, llama-3-8b) and four corpora (Blogs50, CCAT50, Guardian, IMDB62). For each corpus, we extract a smaller sub-corpus and estimate the RF of all types in the full corpus. We find LERF outperforms all other methods, especially when the size of the sub-corpus is limited.

List of references

Gale, W. A., & Church, K. W. (1994). What's wrong with adding one. Corpus-based research into language: In honour of Jan Aarts, 189-200.

McEnery, T., & Hardie, A. (2011). Corpus linguistics: Method, theory and practice. Cambridge University Press.

Valency class complexity and construction of meaning

Andrea Hudousková

Charles University, Faculty of Arts

Czech is considered to be a language with a high degree of valency class complexity (Say, 2024), which is a challenge for L2 learners. While some valency classes are represented by a large number of lemmas, others are quite limited. Furthermore, the lemmas differ in their frequency and also in the proportion of the total number of tokens they represent within a valency class.

Based on the syntactically annotated representative corpus of Czech SYN2020 (Křen et al., 2020), the study examines the frequency of individual valency classes and the type-token relations within them. Given the assumption of construction grammar that frequent lemmas are important for the generalisation of meaning (Goldberg 2006; Herbst 2020), the aim of the study is to focus on how the lemmas in different valency classes contribute to the construction of prototypical meanings of Czech cases and prepositions postulated by Janda&Clancy (2006). The process of entrenchment should be facilitated in valency classes with dozens of frequent lemmas, whereas less frequent valency classes with the majority of tokens represented by a few lemmas are likely to be more problematic for language acquisition.

For example, the valency class of ditransitive verbs is represented by many frequent verbs that provide a sufficient input for generalising the possible meanings of this construction (99 lemmas with ipm>900). On the other hand, 50% of the tokens within the valency class of verbs complemented with the preposition *za* + accusative are represented by the first five lemmas with much lower frequency (ipm≤176).

Such observations shed light on the salience of individual lemmas within a valency class and the proportion of all tokens they represent, and separate classes of verbs that are liable to generalisation from idiosyncratic cases, which has direct implications for L2 teaching and learning.

List of references

References

Goldberg, A.G., 2006.Constructions at work: the nature of generalization in language, New York: Oxford University Press.

Herbst, T., 2020. Constructions, generalizations, and the unpredictability of language: Moving towards colloconstruction grammar.Constructions and frames, 12(1), pp.56-95.

Janda, L.A. & Clancy, S.J., 2006.The Case Book for Czech, Bloomington: Slavica Publ.

Křen, M., Cvrček, V., Henyš, J., Hnátková, M., Jelínek, T., Kocek, J., Kováříková, D., Křivan, J., Milička, J., Petkevič, V., Procházka, P., Skoumalová, H., Šindlerová, J., Škrabal, M., 2020. SYN2020: reprezentativní korpus psané češtiny. Praha: Ústav Českého národního korpusu FF UK. <http://www.korpus.cz>;

Say, S., 2024. Valency Patterns across Slavic: Type vs. Token Frequency. Paper presented at the conference „Valency Research in Slavic Countries – Past and Present“, Prague, October 24-25, 2024.

Picturing the language of patient expertise: A corpus-driven exploration across chronic illnesses

Julie Humbert-Droz[1], **Aurélie Picton**[2]

[1]Sorbonne Nouvelle University, CLESTHIA EA 7345; [2]TIM-FTI, University of Geneva

Patients with chronic illnesses develop considerable knowledge about their conditions and the management of their symptoms over time. Such knowledge is sometimes referred to as *patient expertise*, defined as all the knowledge acquired over the years through patients' lived experiences, their interactions with the healthcare system, and their relationship with medical professionals (Sanderson, Angouri 2013, Flora 2015, Tomasky 2023). From a lexical perspective, this expertise is revealed through patients' appropriation of the terminology typically used by medical experts, be it specific to one illness or to the medical field in general (Fage-Butler, Niesbeth Jensen 2016, Bellander, Landqvist 2020, Delavigne 2021).

Drawing on an experiment reported in Drouin et al. (2018), in the domain of the environment, we seek to identify a shared layer of lexicon among patients with different chronic illnesses that would be typical of patient expertise. To do so, different term and keyword extraction methods (Scott 1997, Drouin 2003, Egbert, Biber 2019) are applied to a 3.2-million-word comparable corpus of texts from French medical experts and patient forums on chronic illnesses (e.g., endometriosis, type 2 diabetes). The results are combined to uncover the units that are specific to both patients' online discourses and the medical field. This approach highlights three main categories of units: 1) health-related terms (e.g., body parts, symptoms, providers), 2) expression/request of support, and 3) daily chronic illness management. Building on these findings, a contextual analysis of these units is performed in the forum subcorpus, which provide further insight into their use by patients and their differences from medical discourse.

These observations help better depict patient expertise from a lexical perspective, thus refining our understanding of how patients appropriate terms and concepts related to chronic illnesses. This supports the enhancement of patient-provider communication and the creation of better-tailored resources for patients.

List of references

Bellander, T., & Landqvist, M. (2020). Becoming the expert constructing health knowledge in epistemic communities online. Information, Communication & Society, 23(4), 507 522.

Delavigne, V. (2021). Phraséologie et didacticité dans les discours de vulgarisation médicale : Une ergonomie discursive. PHRASIS | Rivista di studi fraseologici e paremiologici, 5, 108 129.

Drouin, P. (2003). Term Extraction Using Non-Technical Corpora as a Point of Leverage. Terminology, 9(1), 99 117.

Drouin, P., L'Homme, M.-C., & Robichaud, B. (2018). Lexical Profiling of Environmental Corpora. 11th edition of the Language Resources and Evaluation Conference (LREC 2018), 3419 3425.

Egbert, J., & Biber, D. (2019). Incorporating Text Dispersion into Keyword Analyses. Corpora, 14(1), 77 104.

Fage-Butler, A. M., & Nisbeth Jensen, M. (2016). Medical terminology in online patient–patient communication: Evidence of high health literacy? Health Expectations, 19(3), 643 653.

Flora, L. (2015). Du patient « passif » au patient expert. In C. Déchamp-Le Roux & F. Rafael (Éds.), Santé mentale : Guérison et rétablissement (p. 109 119). John Libbey Eurotext.

Sanderson, T., & Angouri, J. (2013). 'I'm an expert in me and I know what I can cope with': Patient expertise in rheumatoid arthritis. Communication & Medicine, 10(3), 249 261.

Scott, M. (1997). PC Analysis of Key Words—And Key Key Words. System, 25(2), 233 245.

Tomasky, S. I. (2023). Patient Expertise & Self-Management. Master's thesis. The University of Texas at Austin.

Reconciling corpus description and linguistic theory: Pattern, construction, system

Susan Hunston

University of Birmingham

The paper demonstrates how a descriptive corpus study may contribute to theoretical positions. The corpus study is the Cobuild Pattern Grammar project (Francis et al. 1996; Hunston & Francis 2000); the theories are Construction Grammar (Goldberg 2006) and Systemic-Functional Grammar (SFG) (Halliday & Matthiessen 2014). The paper answers two research questions:

1. How can verb complementation patterns be reinterpreted as verb argument constructions (VACs)?
2. How can VACs be used to populate system networks based on specified semantic fields?

The starting point to the current project is corpus lexicography, and 50 complementation patterns used to annotate verb senses in the Cobuild dictionaries of English since 1995. Francis et al. (1996) divided the verbs annotated with each pattern into groups based on meaning. Starting from those meaning groups, over 800 VACs have been identified and described on a web-accessed database (see References). In this paper the process is illustrated with the pattern **V n *to* n** and its 42 proposed constructions.

The next stage in the project is to identify constructions that express a specific semantic field. The study works with 9 semantic fields. For each field, the relevant constructions have been arranged into networks (adopted from Systemic-Functional Linguistics), thus categorising the verb resources in English used to express the field. Each field is expressed as a Meaning Network, or taxonomy, and a more streamlined Systemic Network that highlights distinctive features (Matthiessen 2023). Networks for all 9 fields are available on the website. In this paper the process is illustrated with the semantic field Creation.

The paper illustrates the contribution that detailed corpus research can make to theory. It introduces a comprehensive set of VACs, based on the annotation of every verb in a dictionary of English. It argues for a place for corpus-based construction grammar in SFG, modelling the lexical end of the lexicogrammar continuum.

List of references

Francis G., Hunston S. and Manning E. 1996. Collins Cobuild Grammar Patterns 1: Verbs. London: HarperCollins.

Goldberg A. 2006. Constructions at Work: The Nature of Generalization in Language. Oxford: Oxford University Press.

Halliday M.A.K. and Matthiessen C. 2014. Introduction to Functional Grammar. 4th edition. London: Routledge.

Hunston S and Francis G. 2000. Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English. Amsterdam: Benjamins.

Matthiessen C. 2023. System in Systemic Functional Linguistics. Sheffield: Equinox.

Website: transitivity-net.bham.ac.uk

Examining discourses of hopefulness in mental health recovery narratives

Daniel Hunt

University of Nottingham

First-person accounts of recovery from mental illness are central to contemporary recovery-oriented mental healthcare. Recent evidence shows that reading the recovery narratives of others is therapeutically beneficial for sufferers (Slade et al., 2024). This therapeutic effect hinges upon narratives' potential to generate feelings of *hopefulness* in readers (Rennick-Egglestone et al., 2019). However, little is known about the linguistic characteristics of narratives that foster hopefulness, and hence which texts hold the greatest therapeutic potential for people experiencing psychological distress.

In response, I present findings from a corpus-based discourse analysis of the NEON Collection, a 700,000-word corpus of 486 carefully compiled recovery narratives (https://www.researchintorecovery.com/research/neon/). During two randomised controlled trials, participants (*n*=1,023) accessing the NEON Collection rated its narratives for how hopeful they made them feel, resulting in 3,252 individual ratings. This unique combination of corpus data and user ratings presents a valuable opportunity for a corpus-based discourse study of how features of narrative discourse affect us.

Drawing initially on keyness-based techniques, I illustrate the linguistic and semantic features of narratives that received low, middle, and high hopefulness scores when compared against narratives with contrasting ratings (Baker et al., 2019). Subsequent qualitative analyses demonstrate that readers' feelings of hopefulness reflect both aspects of narrative content – the construction of narrators' agency, inclination towards medicalisation, and the (surprising) prevalence of negative emotions – and the narrativity of the texts themselves. The results suggest a profile of the discursive features of recovery narratives that best promote hopefulness in sufferers, as well as those that do not.

I conclude by outlining the complex implications of these findings for recovery-oriented practitioners and the next steps for this applied CL project, which include integrating visual analyses of narratives, and developing a model to predict user ratings based on the prevalence of linguistic features (c.f. Collins et al., 2023).

List of references

Baker, P., Brookes, G. and Evans, C. (2019). The Language of Patient Feedback: A corpus linguistic study of online health communication. Abingdon: Routledge.

Collins, L., Brezina, V., Demjén, Z., Semino, E. and Woods, A. (2023). 'Corpus linguistics and clinical psychology: Investigating personification in first-person accounts of voice-hearing', International Journal of Corpus Linguistics, 28(1): 28-59.

Rennick-Egglestone, S., Ramsay, A., McGranahan, R., Llewellyn-Beardsley, J., Hui, A., Pollock, K., Repper, J., Yeo, C., Ng, F., Roe, J., Gillard, S., Thornicroft, G., Booth, S. and Slade, M. (2019). 'The impact of mental health recovery narratives on recipients experiencing mental health problems: qualitative analysis and change model', PloS One, 14(12): e0226201.

Slade, M., Rennick-Egglestone, S., Elliott, R.A., Newby, C., Robinson, C., Gavan, S.P., Paterson, L., Ali, Y., Yeo, C., Glover, T., Pollock, K., Callard, F., Priebe, S., Thornicroft, G., Repper, J., Keppens, J., Smuk, M., Franklin, D., Walcott, R., Harrison, J., Smith, R., Robotham, D., Bradstreet, S., Gillard, S., Cuijpers, P., Farkas, M., Ben Zeev, D., Davidson, L., Kotera, Y., Roe, J., Ng, F. and Llewellyn-Beardsley, J. on behalf of the NEON study group (2024). 'Effectiveness and cost-effectiveness of online recorded recovery narratives in improving quality of life for people with non-psychotic mental health problems: a pragmatic randomized controlled trial', World Psychiatry, 23(1): 101-112.

Write what you know: How women and men write females and males in Young Adult fiction

Sally Hunt, Maria Leedham, Sarah Mukherjee

Open University

Writers are encouraged to write what they know to lend authenticity to their representations. Fictional worlds constructed by female and male authors of Young Adult (YA) Fiction are the focus of this paper, based on our 5-million-word corpus of the 50 best-selling YA books in the UK (2017-2022) (Leedham et al. 2024). We identify patterns in representation within the sub-corpora, to see the extent to which they reflect dominant discourses of female and male preoccupations. As with children's fiction, these representations are important due to the impact they may have on young readers' developing identities (Hunt 2025).

Our corpus reflects the domination of Young Adult fiction by female authors: DoRA F (female authors) comprises 34 books and DoRA M 16.

Within each sub-corpus, patterns in the 60 strongest collocates of *she* and *he* (AntConc, Anthony 2014) reveal differences in how female and male authors construct their fictional worlds. Interpreted using concordance lines, in F, emotion collocates total 28% for *she* and 18% for *he*: female emotional words reflect stereotyped notions of femininity like coquettish demeanour or powerlessness (*flounces, purr*, coy and apologizes*), while male emotions indicate more ill will towards others, (*loathes* and *mocks*). In M, 35% of collocates with *she* and 26% of collocates for *he* relate to emotion, with *she* linked to representations of *crying* and *blushing*.

In F, the collocates for *she* and *he* are equally agentive, although themes differ, with female characters performing physically smaller actions. In M, the agency attributed to characters is strongly gendered, with male characters portrayed as agentive three times more often than female characters, and using weapons and being more violent.

These patterns suggest that while female authors create gendered worlds in terms of emotional representation, it is the male authors who tend to polarise the portrayal of both emotions and the characters' impact on the world.

List of references

Anthony, L. (2022) AntConc Version 4.0.10 [Computer Software]

Hunt, Sally (2025) Linguistic representations of Gender in Children's Literature: Feeling, Speaking, Doing. Palgrave Macmillan.

Leedham, M., Hunt S., and Mukherjee, S.J. (2024) Uncovering Discourses of Representation in Young Adult Fiction (DoRA). The Open University. https://wels.open.ac.uk/research/projects/young-adult-fiction-project. Accessed 06 January 2025.

(Semi-)Automating Appraisal Annotation: Using corpus and computational methods to develop a scalable approach to stance analysis in deviant online communities

Madison Hunter, Nicci MacLeod, Ralph Morton, Phil Weber

Aston University

The Appraisal framework (Martin & White, 2005) is a well-established and detailed model for identifying and analysing patterns of stance taking, that is, how we express our feelings/judgments about ourselves, others, and our world. It has been applied in a range of communicative contexts, and in recent years its use has become increasingly common within forensic linguistic research (see Gales, 2020). The larger project from which this paper stems demonstrated Appraisal's practical utility in intelligence analysis, providing valuable insights into stance and dynamics in interactions between users in online communities. However, while Appraisal provides immensely detailed accounts of stance resources, it is a very complex framework, the application of which requires time consuming manual analysis. This limits the size of the datasets to which it can be applied and can limit its appeal as an approach for academics and intelligence analysts alike, particularly when compared with faster (albeit less informative and transparent) computerised methods.

This paper presents an attempt to address Appraisal's scalability issue through the corpus-assisted development of a code to annotate Appraisal features computationally. This would enable the detailed analysis of stance taking while significantly reducing the initial time required for annotation. We outline the corpus approaches taken to identifying and distinguishing between a subset of higher-level Appraisal features, and the degree to which it was possible to dig down into the detail of the framework. We then discuss how these corpus results were used to guide the development of a code to (semi-) automate the process, and compare its performance with that of large language models. The paper focuses on the findings of the study, as well as discussing next steps, areas for potential improvement, and further development.

List of references

Gales, T. (2020). "Prison has been a proper punishment": Investigating stance in forensic and legal contexts. In M. Coulthard, A. May, & R. Sousa-Silva (Eds.), The Routledge Handbook of Forensic Linguistics (2nd ed., pp. 675–693). Routledge.

Martin, J. R., & White, P. R. R. (2005). The Language of Evaluation: Appraisal in English. Palgrave MacMillan.

Tracing roots, shaping futures: A corpus-based exploration of historical narratives and cultural symbols in the Hong Kong diaspora

Wai-Fung Iu

Monash University

This study examines the reconstruction of Hong Kong diasporic identity through the lens of collective memory, focusing on the discourse of Hong Kong Diaspora Organizations (DOs) in the UK, Australia, and Canada. In the aftermath of the 2019 Anti-Extradition Law Amendment Bill (Anti-ELAB) movement and the imposition of the National Security Law (NSL) in 2020, a wave of migration has redefined how overseas Hongkongers connect with their heritage and assert their identity. This research explores how DOs, through their discourse on cultural traditions, rituals, and shared practices, differentiate a distinct Hongkonger identity from broader Chinese cultural frameworks, even within shared customs and festivals.

Using Corpus-assisted Discourse Studies (CADS), the study analyzes DOs' online communication on platforms such as Facebook, focusing on the linguistic patterns and framing strategies used to emphasize Hong Kong's unique cultural ethos. Special attention is given to how DOs represent cultural activities such as the Lunar New Year, Mid-Autumn Festival, and other rituals, highlighting the nuanced ways these practices are reinterpreted to convey values like freedom, democracy, and solidarity. The findings reveal how language and discourse shape collective memory, transforming shared traditions into narratives of resistance and identity preservation.

Findings suggest that DOs strategically mobilize historical and cultural references to assert Hongkongers' distinctiveness. These efforts reflect a broader desire to sustain a diasporic identity that not only preserves heritage but also actively resists cultural assimilation. By examining these linguistic patterns, this research sheds light on the evolving role of cultural memory and tradition in transnational identity construction.

This study contributes to discussions on diaspora studies and the politics of memory by offering a fresh perspective on how Hong Kong DOs utilize discourse to sustain collective memory, adapt traditions, and forge a distinct identity in transnational spaces.

List of references

Alexander, C. (2013). Contested memories: The Shahid Minar and the struggle for diasporic space. Ethnic and Racial Studies, 36(4), 590-610. https://doi.org/10.1080/01419870.2012.674542

Anderson, B. (2016). Imagined communities: Reflections on the origin and spread of nationalism (Revised ed.). Verso.

Appadurai, A. (1996). Modernity at large: Cultural dimensions of globalization (1st ed.). University of Minnesota Press.

Basch, L., Glick-Schiller, N., & Blanc, C. S. (1994). Nations Unbound: Transnational Projects, Postcolonial Predicaments, and Deterritorialized Nation-States. Routledge. https://doi.org/10.1201/9781003071266

Bhabha, H. K. (1994). The location of culture. Routledge.

Brah, A. (1996). Cartographies of diaspora: Contesting identities. Routledge. https://doi.org/10.4324/9780203974919

Chan, M., Chen, H.-T., & Lee, F. L. F. (2017). Examining the roles of mobile and social media in political participation: A cross-national analysis of three Asian societies using a communication mediation approach. New Media & Society, 19(12), 2003-2021. https://doi.org/10.1177/1461444816653190

Chow, S.-l., Fu, K.-w., & Ng, Y.-L. (2020). Development of the Hong Kong Identity Scale: Differentiation between Hong Kong 'Locals' and Mainland Chinese in Cultural and Civic Domains. The Journal of Contemporary China, 29(124), 568-584. https://doi.org/10.1080/10670564.2019.1677365

Cohen, R. (1997). Global diasporas: An introduction. University of Washington Press.

Fong, B. C. H. (2022). Diaspora formation and mobilisation: The emerging Hong Kong diaspora in the anti-extradition bill movement. Nations and Nationalism, 28(3), 1061-1079. https://doi.org/10.1111/nana.12804

Glick-Schiller, N., & Fouron, G. E. (2001). Georges woke up laughing: Long-distance nationalism and the search for home. Duke University Press.

Lee, F. (2020). Solidarity in the Anti-Extradition Bill movement in Hong Kong. Critical Asian Studies, 52(1), 18-32. https://doi.org/10.1080/14672715.2020.1700629

Mathews, G. (1997). Hèunggóngyàhn: On the past, present, and future of Hong Kong identity. Bulletin of Concerned Asian Scholars, 29(3), 3-13. https://doi.org/10.1080/14672715.1997.10413089

Veg, S. (2017). The Rise of "Localism" and Civic Identity in Post-handover Hong Kong: Questioning the Chinese Nation-state. The China Quarterly, 230, 323-347. https://doi.org/10.1017/S0305741017000571

Triangulating AI, topic modelling, and CADS: What can they do for us?

Sylvia Jaworska[1], **Mathew Gillings**[2]

[1]University of Reading; [2]Vienna University of Economics and Business

Varieties of machine learning, for example topic modelling and GenAI, are increasingly used for analyses of language and discourse, areas traditionally explored in (corpus) linguistics. Thus, there is an urgent need to assess what these methods can contribute to our understanding of texts and meanings. Do these 'new kids on the block' pose a 'threat' to our methods or can they complement our analyses? Following the approach of triangulation in CL (see e.g., Marchi and Taylor, 2009; Baker and Egbert, 2018; Gillings and Jaworska, 2025), the current paper conducts a three-way methodological triangulation using LLMs, topic modelling and CADS to identify topics in a 200,000-word corpus of corporate sustainability reports. We focus on topics because topic identification is often the first step to interpreting discourses in texts.

Our three methods are:

A: ChatGPT 4o, used to identify 10 topics within the corpus, and 10 key words within each topic. Analyst uses ChatGPT as a research assistant to interrogate the 10 key themes.

B: Mallet, used to produce a topic model consisting of 10 topics made up of 10 words. Analyst examines documents with the highest proportion of each topic to identify 10 key themes.

C: Sketch Engine, used to produce a keyword list. Analyst examines the top 100 keywords, and their concordance lines, to identify 10 key themes.

Results suggest ChatGPT outputs produce the 'what' of discourse, whereas CADS allows the analyst to find both the 'what' and 'how', including language markers of ideological positioning. Machine learning algorithms, whilst time-efficient, often fail to identify language nuances vital to meaning creation. However, LLM analyses are still in their infancy; new techniques (such as chain-of-thought prompting) are still under development. We aim to open up a space for debate about the role of LLM triangulation in traditionally 'glass box' corpus linguistics.

List of references

Baker, P. and Egbert, J. (eds.). (2016). Triangulating Methodological Approaches in Corpus-Linguistic Research. Routledge.

Gillings, M. and Jaworska, S. (2025). How humans and machines identify discourse topics: a methodological triangulation. Applied Corpus Linguistics.

Marchi, A. and Taylor, C. (2009). If on a winter's night two researchers… a challenge to assumptions of soundness of interpretation. Critical Approaches to Discourse Analysis Across Disciplines, 3(1): 1–20.

Does language matter in business & management scholarship? A CL-based reality check

Sylvia Jaworska[1], **Gerlinde Mautner**[2]

[1]Reading University; [2]Vienna University of Economics and Business

Since the early 20th century, the social sciences have undergone a 'linguistic turn', leading to an increased interest in the role of language in constructing social reality. Or such is the claim. But does this entail an increased interest in the fine-grained analysis of language that linguistics has to offer? As linguists, can we rely on our colleagues from business schools to routinely see research questions through a language lens? Anecdotally, the answer is no.

In this paper, we aim to evidence this hunch with a corpus-based investigation into how business & management scholarship has implemented the linguistic turn. We examine the extent to which language is considered, and what kind of conceptual and methodological tools from linguistics are adopted and how. To this end, we compiled a corpus of articles, totalling more than 6 million tokens, from five major US American and British management journals. Articles were included if they featured one or more of the following terms in their title, abstract or keywords: language, linguistic*, discourse and communication. We used Sketch Engine to study frequencies, collocations and concordances of our terms.

Our results suggest strongly that despite the linguistic turn, language and linguistics still play only a minor role in business & management research, with only a small selection of concepts and methods applied. For example, ‘discourse’ is a more popular concept than ‘language’ and is used with more varied collocations. The bad news for the CL community is that corpus linguistics is underrepresented – despite its demonstrated usefulness for studying business discourse. On the basis of our evidence, we discuss this hitherto untapped potential both epistemologically and in terms of lost opportunities for interdisciplinary research and societal impact.

Cracking the shell: A corpus-based analysis of the functions of shell nouns in American legal contracts

Taft Julian, **Brett Hashimoto**

Brigham Young University

The forms, meanings, and functions of shell nouns (SN), semantically abstract nouns that can package and characterize complex segments of discourse, have been documented in academic discourse and L2 English writing (e.g., Aktas & Cortes, 2008; Benitez-Castro & Thompson, 2015). The abstract nature of SNs lends itself to use where ambiguous language might be favored to allow for a variety of contingencies (Benitez-Castro, 2015). Thus, SNs have also been found to be present in legal language where some of these functions are desirable (Ren et al., 2019). The present study examines the use of SNs in a corpus of 2,700 US legal contracts (53+ million words), spanning 54 contract types. SNs were iteratively identified via lists of shell nouns compiled from previous studies on the syntactic frames that mark their use. Once identified, concordances containing shell nouns were randomly subsampled and categorized according to their meanings and functions using double-coded thematic analysis (Clarke & Braun, 2017). The findings illuminate the strategic use of SNs in contract drafting, offering insights into how they shape the interpretation and execution of contractual obligations. Preliminary analysis indicates that SNs are highly prevalent, perhaps even more than in academic publications because of their utility in abstraction, underspecification, and encapsulation. However, there are uses of shell nouns in contracts that may not have been not documented in previous SN research, for instance, exact repetition of long lists of SN. This work contributes to the understanding of vagueness in legal language, potentially informing both jurists and linguists about the meanings, functions, and interpretation of SNs in contract law. The results of this study could also provide guidance for contract drafting practices as some shell nouns appear to be highly functionally useful/necessary whereas others appear to be merely conventional/superfluous, which may improve clarity and fairness in legal agreements.

List of references

Aktas, R. N., & Cortes, V. (2008). Shell nouns as cohesive devices in published and ESL student writing. Journal of English for Academic Purposes, 7(1), 3-14.

Benitez-Castro, M. A. (2015). Coming to grips with shell-nounhood: A critical review of insights into the meaning, function and form of shell-noun phrases. Australian Journal of Linguistics, 35(2), 168-194.

Benitez-Castro, M. A., & Thompson, P. (2015). Shell-nounhood in academic discourse: A critical state-of-the-art review. International Journal of Corpus Linguistics, 20(3), 378-404.

Clarke, V., & Braun, V. (2017). Thematic analysis. The Journal of Positive Psychology, 12(3), 297-298.

Ren, H., Wood, M., Cunningham, C. D., Abbady, N., Romer, U., Kuhn, H., & Egbert, J. (2019). 'Questions Involving National Peace and Harmony'or'Injured Plaintiff Litigation'? The Original Meaning of 'Cases' in Article III of the Constitution. Georgia State University Law Review, 36, 535.

"[O]f course he'd be confused with all the languages flying about" - A corpus-assisted critical discourse analysis of multilingual parenting discussions on Reddit

Ursula Kania, Sofia Lampropoulou, Paige Johnson, Annarita Magliacane

University of Liverpool

Reddit is a social media platform with over 100,000 active communities and 57 million daily users (Reddit Press, 2021). Registered users ('redditors') post content in so-called 'subreddits' and upvote/downvote contributions made by others. Although Reddit has received some academic attention in relation to controversial communities (e.g., see Chang, 2020 on incels), subreddits containing less hostile content are still underexplored.

Taking a corpus-assisted critical discourse approach (Baker & McEnery, 2015), this study involves the analysis of 13 posts and 3,769 associated comments on multilingual parenting, taken from the AITA (='Am I the Asshole?') subreddit, totalling 252,048 words. AITA is a forum for users to ask questions and discuss moral dilemmas (e.g., 'AITA for refusing to make my bilingual daughter use an English term for my Brother?').

The dataset was extracted using RedditExtractoR package for R (Rivera, 2022). Quantitative analyses (keywords, collocations) were combined with in-depth qualitative coding and analysis for dominant themes/discourses (Fairclough, 2010), using AntConc (Anthony, 2024) and NVivo, respectively. A particular focus was on the exploration of the stances users take vis-à-vis language learning and bilingual/multilingual identities, evaluating which (language) ideologies are reproduced and/or challenged.

We identified three prevalent themes: 1. Family/cultural heritage, 2. Cognitive consequences of bi/multilingualism, and 3. Prestige/stigma. Within all themes, there is a tendency for users to attribute negative evaluations of bi-/multilingualism to others (outside the subreddit), while themselves adopting a positive stance. For example, users may encourage others to pass on their heritage language(s) to their child(ren) while acknowledging the challenges posed by raciolinguistic ideologies, within which the bilingualism of racialised speakers is viewed negatively (Rosa and Flores, 2017).

Overall, the results of this study contribute to and expand on existing research on social media discourse by providing a novel perspective on language ideologies and stances towards multilingualism in online discussion fora.

List of references

Anthony, L. (2024). AntConc (Version 4.2.4) [Computer software]. Retrieved from https://www.laurenceanthony.net/software

Baker, P. and McEnery, T. (eds.) (2015). Corpora and Discourse: Integrating Discourse and Corpora. London: Palgrave.

Chang, W. (2020). "The monstrous-feminine in the incel imagination: Investigating the representation of women as 'femoids' on /r/braincels," Feminist Media Studies, 22(2), pp. 254–270. Available at: https://doi.org/10.1080/14680777.2020.1804976.

Fairclough, N. (2010). Critical Discourse Analysis: The Critical Study of Language. 2nd edition. New York: Longman.

Rivera, I. (2022). 'Package 'RedditExtractoR'. Available at: RedditExtractoR: Reddit Data Extraction Toolkit (r-project.org).

Rosa, J. & Flores, N. (2017). "Unsettling race and language: Toward a raciolinguistic perspective." Language in Society 46, no. 5, pp. 621-647.

Do words know they belong to a register repertoire? – Capturing register variation in contextual embeddings

Antti Olavi Kanner, Erik Henriksson, Maria Veronika Laippala

University of Turku

The introduction of transformers (Vaswani et al. 2017) in recent years has revolutionized language modeling. By representing words as contextual embeddings – word embeddings that model each occurrence of a word with its own unique vector based on the occurrence context – these models are able to capture and represent many kinds of characteristics from linguistic expressions' occurrences, both linguistic and extra-linguistic. In this paper, we examine whether register features (Biber 2012, Biber & Conrad 2019) are such characteristics.

We study the relation between registers and their repertoires from a computational perspective. This research will enable new possibilities in the development of corpus linguistic methods, especially for analysing linguistic structures' distribution across registers and the makeup of registers' repertoires. According to Biber & Conrad (2019), any linguistic expression can be associated with a register and belong to its repertoire. Thus, at least some associations between registers and linguistic expressions should leave a trace which contextual embeddings learn. If so, contextual word embeddings would be able to cluster occurrences following register lines by identifying register markers in the contexts. We test this hypothesis with a large amount of data from the HPLT 2.0 (de Gilbert et al 2024) web dataset, register annotated using the 25-class CORE scheme (Henriksson et al. 2024), and targeting English, Finnish, French and Thai.

In our analysis, we will test how well the embeddings of a test word list are clustered according to their registers. The list contains both randomized sample and register keywords from our data. The words in the list are sampled to represent all grammatical categories and all frequency bands. We compare two language models: one trained for automatic register classification (Henriksson et al. 2024), and another that performs well in clustering but has not been trained on register annotations (Wang et al. 2024).

List of references

Biber, D. (2012). Register as a predictor of linguistic variation. Corpus Linguistics and Linguistic Theory, 8(1), 9–37. https://doi.org/10.1515/cllt-2012-0002

Biber, D., & Conrad, S. (2019). Register, Genre, and Style (2nd ed.). Cambridge University Press. https://doi.org/10.1017/9781108686136

de Gibert, O., Nail, G., Arefyev, N., Bañón, M., van der Linde, J., Ji, S., Zaragoza-Bernabeu, J., Aulamo, M., Ramírez-Sánchez, G., Kutuzov, A., Pyysalo, S., Oepen, S., & Tiedemann, J. (2024). A New Massive Multilingual Dataset for High-Performance Language Technologies. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) (pp. 1116–1128). ELRA and ICCL. https://aclanthology.org/2024.lrec-main.100/

Henriksson, E., Myntti, A., Eskelinen, A., Erten-Johansson, S., Hellström, S. and Laippala, V. 2024. Automatic register identification for the open web using multilingual deep learning. arXiv preprint arXiv:2406.19892.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J. Jones, L., Gomez, A. N., Kaiser, Ł and Polosukhin, I. 2017. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 6000–6010.

Wang, L., Yang, N., Huang, X., Yang, L.,Majumder, R. and Wei, F. 2024 Multilingual E5 Text Embeddings: A Technical Report'. arXiv:2402.05672. https://doi.org/10.48550/arXiv.2402.05672.

Corpus linguistic approaches to safeguarding studies: A case study on suicide-related language

Charlotte-Rose Kennedy[1], **Mark McGlashan**[2]

[1]Birmingham City University/Senso.cloud; [2]University of Liverpool

The need to improve suicide prevention amongst children and young people 'could not be higher' (Sleap et al., 2021: 4) - particularly in online contexts, where young people struggling with suicidal ideation can search for information on methods of suicide and communicate suicidal ideas or intent (Rodway et al., 2023). The UK's Online Safety Act (2024) has introduced a new criminal offence for encouraging or assisting serious self-harm, which highlights the significance of the problem.[1]

To protect young people against such harms, the Department for Education (2024: 40) requires UK schools to implement filtering and monitoring software that enables them to "block harmful and inappropriate content without unreasonably impacting teaching and learning". Many filtering and monitoring systems use 'keyword monitoring' to track language use on online devices to identify specific words or phrases (e.g. 'bomb') that correlate with a specific form of risk (e.g. violence). However, this poses some issues; filtering and monitoring software tends only to raise concerns if there is a direct match to a 'keyword', and the 'keywords' themselves are often isolated from their context(s) of use. This can lead to 'false positives', wherein a keyword match raises an automatic safeguarding concern (e.g. 'bomb') even if the use of the keyword was innocuous (e.g. 'bath bomb').

In this paper, we demonstrate how we use corpus linguistics methods to enhance current practice at Senso.cloud, a safeguarding solutions provider. We outline a study of a 1,094,914-word corpus of online testimonies relating to suicide and suicidal ideation. Specifically, keyword, collocation, and concordance analyses are used to derive a variety of linguistic patterns that are used in natural language to express suicidal ideation (e.g. 'am in immense pain') providing findings that can enable more context-sensitive, empirically-based approaches to 'keyword monitoring'.

[1] https://www.gov.uk/government/publications/online-safety-act-explainer/online-safety-act-explainer#the-act-will-tackle-suicide-and-self-harm-content

List of references

Department for Education, 2024. Keeping children safe in education 2024: statutory guidance for schools and colleges. https://assets.publishing.service.gov.uk/media/6650a1967b792ffff71a83e8/Keeping_children_safe_in_education_2024.pdf (accessed 6.18.24).

Rodway, C., Tham, S. G., Richards, N., Ibrahim, S., Turnbull, P., Kapur, N., & Appleby, L. (2023). Online harms? Suicide-related online experience: a UK-wide case series study of young people who die by suicide. Psychological Medicine, 53(10), 4434–4445. https://doi.org/10.1017/S0033291722001258

Sleap, V., Williams, T., Stoianova, S., Odd, D., Gunnell, D., Chitsabesan, P., Irani, T., Rodway, C., Skelton, S., Tranter, S., King, A., McClymont, C., Fonagy, P., & Luyt. (2021). Suicide in Children and Young People: National Child Mortality Database Programme Thematic Report—Data from April 2019 to March 2020. Healthcare Quality Improvement Partnership (HQIP). https://www.ncmd.info/publications/child-suicide-report/

Interpretative categories in discourse analysis: Challenges in annotation and automation

Carina Kiemes

TU Darmstadt

Interpretative categories that do not necessarily appear on the linguistic surface pose challenges for both annotation and automated classification. Factors such as segmentation, interpretation depth, and contextual knowledge of annotators are crucial (Bender et al. 2023).

As part of a research program analyzing controversial political discourses in Germany, this project reflects methodologically on collaborative annotation, including tag set creation, guideline refinement through inter-annotator agreement, and gold-standard development. It further explores the automated classification of these categories and evaluates model performances both quantitatively and qualitatively.

As a practical use case, I will discuss the category of topoi. The topos concept aligns with Toulmin's model of argumentation (1958), where the warrant provides a structural link between premises and conclusions, enabling rhetorically and contextually acceptable arguments. A topos, as conceptualized by Wengeler (2003: 183 f.), refers to a patterned, content- and theme-specific warrant used in argumentative reasoning. As such, topoi function as shared social thought patterns guiding argumentation within specific discourses.

Focusing on the utility topos ("*If an action provides benefit, it should be realized*"), I address challenges in annotating and automating such categories. Using expert annotations, I train LLMs like German BERT base (deepset/gbert-base) and GPT models (OpenAI 2024) to classify topoi in corpora such as German plenary debates (Müller 2022) and media texts (e.g., FAZ, Die Zeit, taz, BILD, NZZ).

Preliminary results for the utility topos are:

- **German BERT base:** Accuracy **0.91**, F1 0.42
- **GPT-4o-2024_08_06 (no fine-tuning):** Accuracy 0.82, F1 0.62
- **GPT-4o-2024-08-06 (fine-tuned):** Accuracy 0.83, F1 0.60
- **GPT-4o-mini-2024-07-18 (fine-tuned):** Accuracy 0.85, F1 **0.68**

Future experiments will include testing Gemini (GoogleAI) and German GPT-2 (Schweter 2020) models, experimenting with prompts, and exploring few-shot learning approaches (as in Ashok/Lipton 2023). By combining qualitative annotations with automation, this project highlights the potential of improving quantitative discourse analysis with qualitative insights.

List of references

Ashok, Dhananjay/Lipton, Zachary C. (2023): PromptNER: Prompting For Named Entity Recognition. arXiv. https://arxiv.org/abs/2305.15444.

Bender, Michael/Becker, Maria/Kiemes, Carina/Müller, Marcus (2023): Category Development at the Interface of Interpretive Pragmalinguistic Annotation and Machine Learning – Annotation, Detection and Classification of linguistic routines of discourse referencing in political debates. In: Digital Humanities Quarterly Special Issue Working on and with Categories for Text Analysis: Challenges and Findings from and for Digital Humanities Practices, 17 (3). http://www.digitalhumanities.org/dhq/vol/17/3/000720/000720.html.

Chan, Branden/Schweter, Stefan/Möller, Timo (2020): German's Next Language Model. arXiv. http://arxiv.org/abs/2010.10906.

GoogleAI (no year): Gemini [Software]. Retrieved from https://ai.google.dev.

Müller, Marcus (2022): Die Plenarprotokolles des Deutschen Bundestags auf Discourse Lab. In: Korpora Deutsch als Fremdsprache, 2 (1), 123–127. doi: https://doi.org/10.48694/kordaf-3492.

OpenAI (2024): ChatGPT (Version 4o) [Software]. Retrieved from https://openai.com/.

Schweter, Stefan (2020): German GPT-2 Larger [Transformer based language model]. Available at: https://huggingface.co/stefan-it/german-gpt2-larger.

Toulmin, Stephen (1958): The Uses of Argument. Cambridge: Cambridge University Press.

Wengeler, Martin (2003): Topos und Diskurs: Begründung einer argumentationsanalytischen Methode und ihre Anwendung auf den Migrationsdiskurs (1960 - 1985). Tübingen: Niemeyer.

The impact of self-initiated L2 English reading on phraseological variation in student writing

Taehyeong Kim[1], Tove Larsson[1], Henrik Kaatari[2], Ying Wang[3], Pia Sundqvist[4]

[1]Northern Arizona University; [2]University of Gävle; [3]Karlstad University; [4]University of Oslo

Second language (L2) learners worldwide are increasingly exposed to English outside the classroom through self-initiated, Extramural English (EE) activities (Sundqvist, 2009). Previous studies have shown that frequent EE engagement leads to more diverse (Kaatari et al., 2023) and less frequent (Olsson, 2016) vocabulary in L2 writing. EE reading stands out as a particularly important activity for students, as it has been shown to foster L2 learner's knowledge of phrasal verbs (Garnier & Schmitt, 2016), collocations (González-Fernández & Schmitt, 2015), and adjective-noun combinations (Wang et al., 2025). The present study examines whether the effect of EE reading extends to longer multi-word units, specifically to *phrase frames* (*p*-frames): discontinuous multi-word sequences with a variable slot (e.g., *the most * aspect*).

This study uses an EE survey and the Swedish Learner English Corpus (SLEC; Kaatari et al., 2024), which comprises Swedish junior and senior high school writing, to identify learners who read in English every week and those who do not. Based on a novel method to identify *p*-frames that are *key* to the *reading* and *non-reading* groups, our study asks:

- To what extent do the *reading* and *non-reading* groups differ in terms of variability of fillers in key *p*-frames?
- To what extent do the *reading* and *non-reading* groups differ in terms of structural characteristics of key *p*-frames?

The results show that the *reading* group's key *p*-frames are characterized by high variability, a frequent use of post-nominal modifiers (e.g., *the * of a*), and lower frequencies of personal pronouns, i.e., features associated with informational writing (e.g., Biber, 2006). In contrast, the *non-reading* group use key *p*-frames that are characterized by lower variability, embedded clauses and personal pronouns, i.e., features common in more involved production (e.g., Biber, 2006). Pedagogical implications for the role of self-initiated reading in L2 writing development are discussed.

Corpus-assisted discourse studies to the rescue: Unpacking social media research through a four-layered framework

Susanne Kopf

WU Wien

This paper deals with the application of Corpus-Assisted Discourse Studies (CADS) to the analysis of social media. It introduces a four-tiered framework for social media research and demonstrates how each of these layers can be effectively examined using a CADS methodology. In addition, the paper discusses potential weaknesses of the CADS approach and how these can be mitigated.

Social media have yielded ample research in linguistics and discourse studies. Principally consisting of digital language material (additionally to other forms of semiosis), social media data lend themselves to collecting and corpus building. That is, at least in comparison to analogous data which need to be digitised, etc. first, building corpora from social media data is relatively easy and there are even tools that support automated scraping/annotation of social media data[1]. Therefore, it is unsurprising that past research in corpus linguistics has dealt with social media data, ranging from, for example, corpus pragmatic studies on business to customer interaction on Twitter[2] to corpus-assisted discourse analyses of governmental social media communication[3]. Indeed, even the issue of social media's multimodal meaning-making has received research attention in corpus-assisted discourse studies, e.g. Corpus-Assisted Multimodal Discourse Analysis (CAMDA)[4].

This paper complements existing research by introducing a classification scheme of how social media can be studied via CADS: 1) Social media as data repositories, 2) Discourse(s) about social media, 3) Discourse by social media providers, and 4) Discourse analytical contextualisation and theorisation of social media. The presentation explores how CADS can be applied to these four approaches. Here, I address the benefits of taking a CADS approach to social media, e.g. mitigating the danger of cherry-picking data. Additionally, I address potential pitfalls of adopting a CADS perspective and detail how to avoid certain issues when designing a corpus-assisted discourse analytical study of social media.

List of references

Berberich, Kristin, and Ingo Kleiber. 2023. "Corpus Analysis: Tools for Corpus Linguistics." https://corpus-analysis.com/.

Caple, Helen. 2018. "Analysing Multimodal Text." In Corpus Approaches to Discourse: A Critical Review, edited by Charlotte Taylor and Anna Marchi, 85–109: Routledge.

Hansson, Sten, and Ruth Page. 2023. "Legitimation in Government Social Media Communication: The Case of the Brexit Department." Critical Discourse Studies 20 (4): 361–78. https://doi.org/10.1080/17405904.2022.2058971.

Lutzky, Ursula. 2021. The Discourse of Customer Service Tweets: Planes, Trains and Automated Text Analysis. First edition. Bloomsbury Discourse. London, UK, New York, NY: Bloomsbury Academic.

Comparing Wordlists: Do frequency types matter?

František Kovařík[1,2], Vojtěch Kovář[1,3], Marek Blahuš[1], Miloš Jakubíček[1,3]

[1]Lexical Computing CZ, s.r.o.; [2]Faculty of Arts, Masaryk University; [3]Faculty of Informatics, Masaryk University

Building a proper vocabulary is one of the least visible yet most important parts of making a dictionary of the general language. Depending on the preferred type and size of the dictionary, the choice of headwords changes dramatically. The lexicographer should consider several factors, two of which are word frequency and the native speaker's passive knowledge of the words.

Some hypotheses state that the passive vocabulary of a usual native speaker reaches "tens of thousands" of headwords (words of a particular part of speech in their basic form). Yet, few studies in the corpus linguistics field support the subject with more precise numbers or evidence. Our ongoing research focuses on examining the data of Czech corpora, especially comparing frequency types. The first research phase should answer three questions about the wordlists of the most frequent headwords based on the absolute frequency, document frequency, Average Reduced Frequency (ARF) and Average Logarithm Distance Frequency (ALDF) of the headwords and their forms:

1. The percentage of the accepted (i.e. "proper Czech", based on the revised annotation data) headwords from the top 10,000, 50,000 and 100,000 most frequent headwords of each frequency type.

2. The quality of the headwords accepted in the manual annotation, yet missing from the top 100,000 most frequent headwords of each frequency type.

3. The qualitative difference between wordlists of the top 100,000 most frequent headwords of any two frequency types.

This comparison will later be used to study the passive vocabulary of Czech and examine the annotation data of more than 100,000 of the most frequent Czech headwords.

Exploring AI translationese: A machine learning approach to comparing linguistic patterns in AI and human learner translations

Ho Ling Kwok, Yanfang Su, Yiyang Hu, Kanglong Liu

The Hong Kong Polytechnic University

Translations often exhibit distinctive linguistic patterns that set them apart from native texts, a phenomenon referred to as “translationese.” These patterns arise as traces of the translation process. In the context of machine translation (MT), computational algorithms also leave unique imprints on translated outputs, leading to the emergence of “machine translationese.” With the rapid advancement of generative artificial intelligence (GenAI) technologies, such as GPT-4o, there has been growing interest in exploring AI-generated translations and their implications for translation education and professional practice.

Although prior research has extensively examined the quality of AI-generated translations and user perceptions, limited attention has been given to the potential existence of AI translationese and its comparison with human learner translationese. This study seeks to bridge this gap by investigating the linguistic features of AI-generated translations alongside those produced by learner translators, using non-translated native texts as a benchmark. The analysis includes a univariate, contrastive approach to identify individual linguistic features that reveal constraints specific to AI and learner translations. These features are subsequently evaluated using supervised machine learning classifiers to determine their effectiveness in detecting translationese and distinguishing between AI-generated and learner translations.

By integrating corpus linguistics with supervised machine learning techniques, this research uncovers patterns that distinguish AI-generated translations from learner translations and native texts. The findings provide valuable insights into the nature of AI translationese, its overlap and divergence with learner translationese, and its broader implications for the application of GenAI tools in translation practice and education. This study contributes to the understanding of the linguistic limitations and potentials of GenAI technologies, paving the way for their more effective use in the translation industry and academic settings.

"Hong Kong will prosper only when its young people thrive": The discursive construction of Hong Kong in policy addresses

Phoenix Lam, Richard Sandes Forsyth

The Hong Kong Polytechnic University

This study examines the discursive construction of the city of Hong Kong through a sub-genre of political speech known as policy address – an annual speech given by the head of the city to the legislature to share and spread the overarching governance ideologies and specific strategic plans of the government. Generally regarded as one of the most carefully planned and scripted political texts produced in the city, the policy address offers a key window through which policy development of the city can be investigated. Focusing on the three policy addresses delivered by the current Chief Executive, the study employs a corpus-based approach to uncover the lexico-grammatical patterns of the proper noun "Hong Kong". Specifically, it compares the different positions the proper noun occupies in a clause. Then concentrating on "Hong Kong" taking the subject position of the clause, the study identifies the most frequently occurring main verbs, the tense choices of the verbs, and other collocates commonly co-occurring with the city's name. Findings from the study show that Hong Kong is mainly constructed as part of circumstance, rather than theme, in the policy address. It is also most frequently conceptualised as engaging in the relational processes of being and owning, rather than in behavioural or material processes. Its strong association with the present tense as clause subject, in particular, reveals a weak orientation to both the past and the future. Further, its frequent lexical and grammatical collocates suggest two major strategies employed in the discourse of policy address – quantification of government performance and national level intertextual reference. The study concludes by considering the implications of the current governance ideologies and the discursive construction of Hong Kong in the policy address for the future of the city.

Evaluating the extent to which corpus size and inclusion criteria affect ranking stability of word lists using bootstrapping

Kyra Larsen[1], Brett Hashimoto[2]

[1]Northern Arizona University; [2]Brigham Young University

Word lists are most beneficial to learners when they include, and properly rank, the most representative/useful words within a language domain. To create word lists, researchers have used corpora with a wide range of sizes and varying inclusion criteria (Tong et al., 2025). However, this variance between studies raises questions about which procedure produces the most representative/useful list. The present study uses bootstrapping to randomly sample texts from the 2,700 text, 53+ million word Corpus of English Business Contracts (Hanks et al., 2024) to assess list stability by assessing the extent to which different subsamples produce the same list rankings (Egbert & Plonsky, 2021) to explore how corpus size and inclusion criteria (i.e., range, dispersion) impact word list stability. Each random bootstrapped sample of the corpus was first manipulated for total corpus size in four conditions: 1, 2, 10, and 50 million words. Then, different ranking criteria were used (only range, only dispersion, and both) to assess which yielded the most stable results for ranking in each corpus size condition. Through 500 bootstrapped resamplings, confidence intervals for the word rankings in each size + criteria condition were produced. Preliminary results reveal that the stability of the word list decreased sharply (a sharp increase in confidence intervals) after only ~200 words for corpora of smaller sizes (1-2 million words). This result indicates that small corpora may be simply insufficient to achieve a high degree of stability beyond a couple hundred words, even within specialized corpora, consistent with previous research (e.g., Burch & Egbert, 2023). The results also showed that the use of criteria does improve the stability of word lists. These results indicate that larger corpora are needed to create stable word lists than are frequently used and that other inclusion criteria are also necessary beyond frequency alone.

List of references

Burch, B., & Egbert, J. (2023). Confidence intervals for ratios of means applied to corpus-based word frequency classes. Journal of Applied Statistics, 50(7), 1592-1610.

Egbert, J. & Plonsky, L. (2021). Bootstrapping techniques. In M. Paquot & S. Gries (eds.). A practical handbook of corpus linguistics (pp. 593-610). Springer International Publishing.

Tong, J., Che Abdul Rahman, A. N., & Hamat, A. (2025). A systematic literature review on word selection criteria in corpus-based wordlists development. Cogent Arts & Humanities, 12(1), 2432131.

How to detect similarity between groups rather than a difference: Equivalence testing for corpus linguistics

Tove Larsson[1], Gregory R. Hancock[2]

[1]Northern Arizona University; [2]University of Maryland

If we want to assess whether we see similarity between groups in a characteristic of interest, we are out of luck if we use the most established analytical methods in the field. This is because in the null hypothesis significance testing (NHST) framework, the main goal is to identify differences between groups. As most commonly practiced, the null hypothesis states that there is no difference between populations of interest in terms of a specific parameter (e.g., mean, regression slope), whereas the alternative hypothesis states that there is some non-zero difference. We then conduct our statistical test (e.g., t-test, regression), hoping to reject the null hypothesis to make the inference that there is some difference in the populations. However, as we know, if we do not have enough power to reject the null hypothesis, we cannot "prove" the null hypothesis, merely fail to reject it (e.g., Bonovas & Piovani, 2023).

However, in many study designs in corpus linguistics, it is relevant to know whether there is, in fact, similarity between groups. This paper provides a non-technical introduction to methods that can be used to assess practical equivalence. To illustrate, we re-analyze data from Biber et al. (2025) using two techniques: equivalence testing (e.g., Lakens et al., 2018) and mean structure models from the structural equation modeling family (SEM; see, e.g., Larsson et al., 2024). The following research question will guide the example:

Are L1 and L2 English students' frequency of use of the complexity feature attributive adjectives similar enough such that we can merge the two groups into a single 'student group'; that is, are the two groups practically equivalent?

In the presentation, we will go through the steps of carrying out the analysis using these techniques to see whether we can draw the conclusion that the groups are practically equivalent.

List of references

References

Biber, D., Larsson, T., Hancock, G. R., Reppen, R., Staples, S., & Gray, B. (2025). Comparing theory-based models of grammatical complexity in student writing'. International Journal of Learner Corpus Research. Early online access.

Bonovas, S., & Piovani, D. (2023). On p-values and statistical significance. Journal of Clinical Medicine, 12(3), 900.

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. Advances in Methods and Practices in Psychological Science, 1(2), 259-269.

Larsson, T., Biber, D., & Hancock, G. R. (2024). On the role of cumulative knowledge building and specific hypotheses: The case of grammatical complexity, Corpora, 19(3), 263-284.

The frequency and use of Lexical Bundles (LBs) in L2-Vietnamese and L1-English discourse compared to the Academic Formula List (AFL)

Nhu Ngoc Quynh Le

The University of Ostrava

Lexical bundles have attracted significant attention from ESL researchers, concerning their role in improving student writing performance, so various quantitative and qualitative analysis have emerged in this field. Limited contrastive studies investigate the use of LBs in written essays at tertiary levels; hence, this corpus-based study is conducted to identify the most commonly used LBs in English texts written by L1-Vietnamese and L1-English undergraduates, to examine how these LBs differ in structural and functional uses, and to compare how much these LBs align with those listed in the AFL (Simpson-Vlach & Ellis, 2010). To achieve these, firstly, a dataset of two learner corpora, namely an unpublished L2-Vietnamese corpus (55,671 words) and a (sub)-corpus from LOCNESS (Granger, 1998) used as L1-English corpus (52,980 words) was analyzed. Secondly, given that LBs, according to Biber et al. (1999), represent key constitutive elements that influence user writing proficiency, the analysis of structures and functions of the most frequently used LBs was carried out. Thirdly, the AFL is used to compare with two frequency lists. The results indicate that L1-Vietnamese students deploy more LBs than L1-English ones, with total normalized numbers per 10,000 words of 40 vs. 28, and frequencies of 395 vs. 243, respectively. Both groups of students use ‘Noun phrase with of-phrase fragments’, ‘Other prepositional phrases’, and ‘Pronoun/ Noun phrase + be/verb phrases’ as the top three frequent patterns. Regarding discourse functions, both show a preference for using referential bundles. While L1-Vietnamese undergraduates often convey abstract concepts and quantitative attributes, L1-English students prefer using bundles to emphasize and express main points. Moreover, there is some degree of overlapping LBs, indicating that the AFL may include commonly used LBs being relevant in both contexts, regardless of the domains or English proficiency. Consequently, this suggests the AFL may potentially improve student LB competence.

List of references

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). Longman Grammar of Spoken and Written English (1st ed.). Pearson Education Limited.

Granger, S. (1998). The computer learner corpus: A versatile new source of data for SLA research. In Granger, S. (ed.) Learner English on Computer. Addison Wesley Longman: London & New York, 3-18.

Simpson-Vlach, R., & Ellis, N. C. (2010). An academic Formulas List: New Methods in Phraseology Research. Applied Linguistics, 31(4), 487–512. https://doi.org/10.1093/applin/amp058

Coding, classifying and checking corpus data: The contribution of Gen AI

Maria Leedham, **Sarah Mukherjee**, **Sally Hunt**

The Open University

The widespread availability of Generative AI (Gen AI) presents researchers with unprecedented opportunities for free research support (Davison et.al, 2024), with Gen AI frequently used for tasks such as generating initial ideas, providing literature suggestions, coding, and formatting data tables. But what can Gen AI do to help the corpus linguist in particular? And what do researchers need to be mindful of? This presentation draws on our experiences with ChatGPT (OpenAI, 2024) to assist with an ongoing project: exploring a 5-million-word corpus of contemporary Young Adult fiction (Discourses of Representation in Young Adult fiction [DoRA], Leedham et al., 2024). Within DoRA, we have employed ChatGPT with the following specific corpus linguistic tasks: generating potential thematic categories for keywords, placing keywords into a limited range of these categories, and coding a sample of concordance lines as literal or metaphorical usages (cf. Curry et al, 2024). We have additionally employed ChatGPT to classify books into genre categories and to justify these groupings. In doing so, the following practices have emerged as important to successful usage: crafting precise prompts (as instructions are followed literally), checking suggestions (for errors, omitted or 'hallucinated' additional lexis) and treating each set of generated results in isolation. The Gen AI imperative to produce human-like text (cf. Sardinha, 2024) means that categorisations omit outlier keywords and subsequent iterations ignore previously-generated outputs. Gen AI carries out some tasks very well, but the uniquely human ability to bring all contextual data to the fore remains essential. An additional consideration is that any negative tropes within the Gen AI source input are simply reproduced in outputs, leading to the inherent danger of stereotypes being continually recirculated. Echoing Zappavigna (2023), we argue that ultimately Gen AI can support the human researcher in evoking insights, but cannot (yet) replace human criticality or analyses.

List of references

Berber Sardinha, B., (2024), AI-generated vs human-authored texts: A multidimensional comparison, Applied Corpus Linguistics, Volume 4 (1). https://doi.org/10.1016/j.acorp.2023.100083.

Curry, N., Baker, P., & Brookes, G. (2024). Generative AI for corpus approaches to discourse studies: A critical evaluation of ChatGPT. Applied Corpus Linguistics, 4(1), 100082. https://doi.org/10.1016/j.acorp.2023.100082.

Davison, R.M., Chughtai, H., Nielsen, P., Marabelli, M., Iannacci, F., van Offenbeek, M., Tarafdar, M., Trenz, M., Techatassanasoontorn, A.A., Díaz Andrade, A. and Panteli, N. (2024), The ethics of using generative AI for qualitative data analysis. Inf Syst J, 34: 1433-1439. https://doi.org/10.1111/isj.12504

Leedham, M., Hunt S., and Mukherjee, S.J. (2024) Uncovering Discourses of Representation in Young Adult Fiction (DoRA). The Open University. https://wels.open.ac.uk/research/projects/young-adult-fiction-project. Accessed 06 January 2025.

OpenAI. (2024). ChatGPT (Version free). https://chat.openai.com

Zappavigna, M. (2023). Hack your corpus analysis: How AI can assist corpus linguists deal with messy social media data, Applied Corpus Linguistics, Volume 3 (3). https://doi.org/10.1016/j.acorp.2023.100067.

Python-assisted source domain metaphor identification and iterative validation: COVID-19 war and journey metaphors in China and the UK

Danyang Li

Lancaster University

Covid-19 presented communicative challenges for governments who needed to find ways to talk about the pandemic. Metaphor was one key resource used in framing the pandemic (Pfrimer & Barbosa 2020). While previous research has shown the various source domains relied upon in figurative framings of COVID-19, including War and Journey (Semino 2021), few studies have undertaken a cross-cultural analysis to consider variation in metaphor usage according to cultural context.

This paper provides a python-assisted cross-cultural examination of metaphorical framing efforts within COVID-19 discourses. Applying Conceptual Metaphor Theory (Lakoff & Johnson 2008) together with Cultural Dimension Theory (Hofstede 2001) as the cultural comparison framework, the research focuses on the following two questions:

RQ1: Are there differences in metaphor use within the Chinese corpus compared to the English corpus?

RQ2: Can different cultural dimensions explain these differences?

Two corpora were collected representing official communication from Dec. 31st 2019, to Dec. 31st 2021. The English corpus consists of 87 government press briefings of 90,372 words, while the Chinese consists of 48 of 73,325 words. The two corpora were manually and automatically searched using a Python script already written to extract potential target metaphor hits. It scans text files in a specified directory, extracts content sentence by sentence, matches predefined metaphor categories, and outputs the matching results as CSV files. Following and further adapting the list of Wicke and Bolognesi (2020), three public databases will be referred to in compiling the directory: 1) *relatedwords.org*, 2) the MetaNet (Karlberg & Buell 2005: 22-39), and 3) the USAS tagger in Wmatrix.

Results show that the Chinese corpus contains significantly more War metaphors, while the English corpus contains more Journey metaphors. A further finding is that the Chinese corpus includes many more idiomatic metaphors, most of which draw on NATURE domain.

List of references

Hofstede, G. (2001). Culture's Consequences. Comparing Values, Behaviors, Institutions, and Organizations across Nations. 2nd ed. . London: Sage.

Karlberg, M., & Buell, L. (2005). Deconstructing the “War of all against all”: The prevalence and implications of war metaphors and other adversarial news schema in TIME, Newsweek, and Maclean’s. Peace and Conflict Studies, 12(1), 22-39.

Lakoff, G., & Johnson, M. (2008). Metaphors we live by. University of Chicago press.

Pfrimer, M. H., & Barbosa Jr, R. (2020). Brazil’s war on COVID-19: Crisis, not conflict—Doctors, not generals. Dialogues in Human Geography, 10(2), 137-140.

Semino, E. (2021). “Not soldiers but fire-fighters”–metaphors and Covid-19. Health communication, 36(1), 50-58.

Wicke, P., & Bolognesi, M. M. (2020). Framing COVID-19: How we conceptualize and discuss the pandemic on Twitter. PloS one, 15(9), e0240010.

Newspapers representations of Niangpao identity in China: A corpus-assisted critical discourse study

Run Li

Lancaster University

In China, “娘炮 (*niangpao*)” is a term used to disparagingly refer to men who exhibit characteristics commonly associated with stereotypical femininity in terms of appearance, behaviour, or mannerisms. In the Chinese context, terms with meanings similar to *niangpao* also include “娘娘腔 (niangniangqiang)” and “伪娘 (weinang)”. To date, no studies have examined how different terms are used to generally represent the *niangpao* identity within the Chinese context. Accordingly, this study adopts a corpus-assisted discourse analysis approach (Baker et al., 2008), using “娘炮 (niangpao)”, “娘娘腔 (niangniangqiang)”, and “伪娘 (weinang)” as search terms to examine representations of the *niangpao* identity in 24 categories of Chinese official newspapers from 2000 to 2023. By employing the word sketch function in Sketch Engine, the study conducts an in-depth analysis of the collocates and concordance lines of these three terms, investigating their representations through aspects such as semantic prosody, semantic categories, and the broader characteristics of the node words. The findings reveal that all three terms are predominantly negatively evaluated in news discourse, with their negative semantic connotations gradually solidifying, reinforcing the stigmatisation and “othering” of the *niangpao* identity. For the differences, “娘炮 (niangpao)” has the lowest frequency of use and is often associated with the government’s official stance against *niangpao* identities. “娘娘腔 (niangniangqiang)” occurs more frequently, is commonly found in everyday contexts, and has a more colloquial tone. “伪娘 (weinang)” has a more focused meaning, primarily referring to “gender-ambiguous” appearances, emphasising essentialist views of gender by highlighting the inherent differences between male and female appearances. This study would offer deeper insights into the broader representations of *niangpao* identity and contribute to a more comprehensive understanding of masculinity within the Chinese culture.

List of references

Baker, P., Gabrielatos, C., Khosravinik, M., Krzyżanowski, M., Mcenery, T., & 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 & Society, 19(3), 273–306. https://doi.org/https://doi.org/10.1177/0957926508088962

Distorted realities: Gender bias in Chinese media coverage of domestic violence

Scarlett Sijia Li

Lancaster University

Domestic violence is a serious social issue in the world. According to the WHO (2021), one in three women experience domestic violence in their lifetime. In China, this figure was 24.7% based on a national survey in 2010 (Zhang, 2014). Despite the severity of the problem, there is not much research into its linguistic or discursive representations in China. The study aims to combine corpus linguistics with critical discourse analysis (Baker et al. 2008) to investigate discourses on domestic violence in an 11.4-million-word corpus of Chinese news articles from ten broadsheet newspapers and four popular news websites published between 2010 and 2019. This study aims to answer the following two research questions: (i) How are wives and husbands represented in Chinese news reports about domestic violence? (ii) What are the ideological implications for public awareness of domestic violence?

The study found the term domestic violence is predominately described as physical violence, which deepens public ignorance of other severe forms of domestic violence such as psychological, sexual and economic abuse. The study revealed that in reports of domestic homicides, wives are overrepresented as intentional homicide offenders, while husbands' intent in committing homicides is more likely to be downplayed. In reports of non-lethal violence, wives as perpetrators are entertainmentized, while husbands as victims are derogatorily represented. Wives' actual agency in instigating divorce after experiencing domestic violence is reduced by positioning them as co-agents while wives' difficulties in getting divorced through judicial procedures are reported as normal. The study presents new findings in the representations of wives and husbands in domestic violence in Chinese media and contributes to the examination of gender bias and patriarchal ideology in reports of domestic violence.

List of references

Baker, P., Gabrielatos, C., KhosraviNik, M., Krzyżanowski, M., McEnery, T., & 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 & Society, 19(3), 273–306.

World Health Organization. (2021). Violence Against Women Prevalence Estimates 2018. World Health Organization.

Zhang, H. (2013). Domestic violence and its official reactions in China. In The Routledge handbook of Chinese criminology (pp. 224-237). Routledge.

Comparing academic formulas in TED talks and TED-Ed animations: Insights for EAP teaching and learning

Chen-Yu Liu

National Central University

Mastering academic formulaic sequences is essential for learners studying English for academic purposes, as these sequences are ubiquitous in academic discourse and serve important communicative functions. Identifying resources that provide frequent encounters with such sequences is crucial to facilitate learning. TED talks and TED-Ed animations, with their strong association with academic topics and extensive coverage of academic single words, may hold potential as pedagogically useful resources for learning academic formulaic sequences. To investigate this potential, this study analyzes the presence of formulas from the Academic Formulas List (AFL; Simpson-Vlach & Ellis, 2010) in terms of frequency and variety across three corpora: TED talks, TED-Ed animations, and academic lectures. The academic lectures corpus serves as a benchmark to assess the relative pedagogical usefulness of TED talks and TED-Ed animations for learning these sequences. The results reveal that while AFL items occur less frequently in TED talks and TED-Ed animations compared to academic lectures, the variety of AFL items in TED talks is comparable to that in academic lectures, suggesting the potential of TED talks for offering encounters with diverse academic formulas. Furthermore, TED talks exhibit significantly higher frequency and variety of AFL items compared to TED-Ed animations. These findings suggest that TED talks may be pedagogically more useful than TED-Ed animations, given their broader coverage of AFL items. However, TED-Ed animations, despite having a relatively lower presence of AFL items compared to TED talks and academic lectures, demonstrate a much higher presence of AFL items than other learning resources (e.g., textbooks). They may serve as complementary materials to TED talks, supporting learners' learning of academic formulas. This study broadens the range of potential resources available for learning academic formulas and deepens our understanding of the pedagogical value and lexical characteristics of TED talks and TED-Ed animations from a phraseological perspective.

List of references

Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. Applied linguistics, 31(4), 487-512.

Exploring TED Talks as a resource for disciplinary vocabulary learning: A case of computer science

Chen-Yu Liu[1], **Jie-Fan Chang**[2]

[1]National Central University; [2]National Taiwan University

Disciplinary vocabulary knowledge is essential for ESP learners to comprehend and produce language effectively in disciplinary contexts. While disciplinary word lists serve as primary tools for learning such vocabulary, it is equally important to identify resources that enable learners to frequently encounter these words and observe their usage in specialized contexts. This study explores the potential of computer science (CS)-related TED Talks as a resource for developing disciplinary vocabulary knowledge, using a corpus of 206 such talks. A lexical profiling analysis examines the extent to which CS-specific words, as defined by the Computer Science Academic Vocabulary List (CSAVL) (Roesler, 2021), occur in these talks, assessing their potential usefulness for vocabulary development. Additionally, a manual analysis evaluates whether CSAVL words in the talks convey general or disciplinary meanings, determining their effectiveness in exposing learners to specialized vocabulary usage. Results indicate that a wide range of CSAVL words occur frequently in these talks, with three-fourths of the CSAVL word types represented, and approximately one in ten words in the talks being discipline-specific. On average, viewing a single CS-related TED Talk can expose learners to 193 CS-specific words and 75 distinct types. Extensive viewing may promote repeated encounters with these words, potentially facilitating vocabulary learning. Manual analysis reveals significant variation in the meanings of disciplinary words within the talks, suggesting that learners may encounter disciplinary words expressing both general and specialized meanings. However, talks with higher coverage of disciplinary vocabulary tend to include more discipline-specific meanings, suggesting that coverage data may serve as a guide for teachers in selecting appropriate materials. This study highlights the pedagogical potential of TED Talks for ESP vocabulary instruction and offers insights into the lexical characteristics of disciplinary vocabulary.

List of references

Roesler, D. (2021). When a bug is not a bug: an introduction to the computer science academic vocabulary list. Journal of English for Academic Purposes, 54, 101044.

Measuring morphological irregularity without manual segmentation: A similarity-based information-theoretic approach

Hewei Liu

University of Neuchâtel

Morphological irregularity has traditionally been viewed dichotomously (e.g., Pinker & Prince, 1991), but recent research argues for a continuum (e.g., Marzi & Pirrelli, 2023). Quantifying degrees of irregularity enables finer modeling and analysis, yet existing approaches either ignore the implicative structure of paradigms (Smith et al., 2023), or they overlook paradigm-internal consistency (Wu et al., 2019). This study proposes a similarity-based, information-theoretic approach to quantifying irregularity without manual inflectional class assignment or morpheme segmentation. We address three research questions: (1) how to measure morphological irregularity systematically and automatically, (2) how English irregular verbs are distributed along a continuum, and (3) whether frequency or dispersion better predicts irregularity (Divjak & Caldwell-Harris, 2015). Our method involves three steps: (i) LCS-based paradigm summarization, which extracts shared stems using the longest common subsequence (LCS); (ii) similarity-based irregularity computation, which quantifies unpredictability of exponent correspondences across paradigms; and (iii) optional exemplar selection, which groups phonologically similar lexemes to enhance paradigm representation. Applied to 1,561 English verbs from the British National Corpus (BNC Consortium, 2007), our approach reveals that irregularity forms a gradient, with some verbs (*make*, *cost*) exhibiting lower irregularity than others (*go*, *run*). Regression analyses indicate a significant quadratic relationship between frequency and irregularity, while dispersion (measured by KLD, Gries, 2024) has a weak and inconsistent effect on its own. These results support frequency as an important driver of irregularity preservation, reinforcing the entrenchment hypothesis (Haspelmath & Sims, 2010). By offering a gradient-based, segmentation-free method, this study provides an operationalized framework for investigating morphological irregularity as an emergent formal property of words.

List of references

BNC Consortium (2007). The British National Corpus, XML Edition. Oxford Text Archive.

Divjak, D., & Caldwell-Harris, C. L. (2015). Frequency and entrenchment. In Handbook of cognitive linguistics, 39, 53–75.

Gries, S. T. (2024). Frequency, dispersion, association, and keyness: Revising and tupleizing corpus-linguistic measures. John Benjamins.

Haspelmath, M., & Sims, A. (2010). Understanding morphology (2nd ed.). Routledge.

Marzi, C., & Pirrelli, V. (2023). A discriminative information-theoretical analysis of the regularity gradient in inflectional morphology. Morphology, 33(4), 459–509.

Pinker, S., & Prince, A. (1991). Regular and irregular morphology and the psychological status of rules of grammar. In Annual Meeting of the Berkeley Linguistics Society, 230–251.

Smith, K., Ashton, C., & Sims-Williams, H. (2023). The relationship between frequency and irregularity in the evolution of linguistic structure: An experimental study. In Proceedings of the Annual Meeting of the Cognitive Science Society, 45(45).

Wu, S., Cotterell, R., & O'Donnell, T. J. (2019). Morphological irregularity correlates with frequency. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 5117-5126.

Decoding emotional expressions in Folk letters: A discourse study of interpersonal relationships

Li Liu

Zhejiang Gongshang University

This mixed-methods study analyzes 6,493 private letters spanning the 1950s to 2010s to trace the evolution of emotional expression across interpersonal relationships. Utilizing the Affective Lexicon Ontology approach, we systematically classify high-frequency emotional words—including positive ones (joy, blessing, tranquility, approval, love) and negative ones (melancholy, dissatisfaction, suffering, uneasiness)—within correspondence between romantic partners, friends, classates, siblings, and parent-child dyads.

Our findings reveal that certain emotion types (including joy, blessing, tranquility, approval, melancholy, dissatisfaction) demonstrate remarkable stability within their specific relational contexts. Particularly, through collocation and concordance analysis of emotions, we have identified that letters between spouses frequently intertwine the “suffering” of prolonged separation with enduring “love”; parent–child exchanges are dominated by “uneasiness” over labor, work, survival pressure and health; siblings use letters as outlets for expressing “dissatisfaction” with daily life and family dynamics; peer correspondence centers on discussions of marriage and mate selection. The diachronic analysis identifies a significant surge in explicit and direct “love” expressions during the 1990s, marking a change from the restrained linguistic conventions of 1960s-1980s letter writing. This shift toward direct emotional disclosure aligns with broader sociocultural liberalization and the higher acceptance of vernacular intimacy in personal communication.

By treating private letters as “fragmented archives” of societal history, this investigation into the real-life emotional exchanges in folk letters illuminates language’s dual role as both mirror and mediator of human connection. And the observed patterns of emotional flow across different social ties offer empirical insights for contemporary psychological studies on social support networks. Furthermore, despite the rise of digital immediacy in modern times, handwritten letters remain a potent medium for personalized emotional articulation, underscoring language’s enduring power to forge profound interpersonal bonds.

Corpus-based quantification and assessment of translation quality in the context of translation training

Yanmeng Liu

Northwestern Polytechnical University

Translation quality is vital in translation training. In the context of translation training, a quality assessment method is expected to benefit students by providing them with an objective judgement of their translation quality, efficient information about their differences from the professional translations, and constructive suggestions for further improvement. Unfortunately, existing assessment metrics fail to meet the expectations, and in practice, translation quality assessment in the training context still heavily relies on human evaluation. This paper uses corpora to investigate and analyse translation quality of student Chinese-English translations in comparison with those of professionals, and has the objectives to quantitatively explain and measure variation in translation quality, and to identify ways to improve it. The paper populates 1,200 translation samples organized into fiction, magazine and news genres, and collects the data of 34 linguistic features of all the samples. Adopting a data-driven approach, the paper first identifies and selects distinctive patterns in the feature used to distinguish the quality of student translations from those of professional translations with unidimensional analyses. Then, multidimensional analyses reveal five dimensions underlaying the selected features. Factor scores of samples along each dimension explain variation in translation quality against the reference line set by professional translations, which specific suggestions to improve student translation quality can be based on. Finally, a model is proposed for future corpus-based translation quality assessment to follow, which improves objectivity, generalisability, and application of the results in an educational context.

Thematic progressions in L2 Czech and L1 English student academic discourse: Types, text frequency and text distributions

Jiří Lukl

University of Ostrava

Drawing primarily on the Prague school approach to the study of information structure in discourse, this corpus-driven study investigates the use of thematic progressions (TPs) by Czech university students of English (L2) and compares them to the use of thematic progressions by British students (L1). TPs are one of the means through which texts build and maintain their cohesion, which in turn is one of the key aspects of mastering academic writing. Nevertheless, in the Czech context, previous studies on cohesion in L2 English student texts have mainly focused on their Theme zones (Dontcheva-Navrátilová et al., 2020), while their TPs have received little attention.

Methodologically, the paper relies on TP types outlined by Daneš (1974), and expanded on by others (e.g., Hawes, 2015; Ho, 2011). For the analysis, the paper uses two corpora: 1) the LOCNESS corpus (specifically, a subset comprising British university students' argumentative essays; 19,091 words); 2) a corpus of argumentative essays written by third-year university students of English collected at a Czech university (25,985 words). Both corpora were tagged and analysed manually.

The aims of the study are, first, to identify the most frequent TPs (and their combinations and distributions across text) found in both corpora and to ascertain to what degree the two corpora differ in this respect. Secondarily, it aims to identify TP breaks (Hawes, 2015) and their distribution with respect to the TPs. Preliminary results indicate that the British L1 essays prefer TPs with a constant theme or simple TPs, both often in combination with TPs with derived themes. Furthermore, TP breaks in these essays are infrequent and often align with paragraph boundaries. On the other hand, the Czech L2 essays mainly utilize TPs with derived themes combined with constant gapped TPs. Importantly, TP breaks occur frequently and somewhat unpredictably in the Czech corpus.

List of references

Daneš, F. (1974). Functional sentence perspective and the organization of the text. In: F. Daneš (Ed.), Papers on functional sentence perspective. Academia (pp. 106–128), Prague. http://dx.doi.org/10.1515/9783111676524

Dontcheva-Navrátilová, O., Jančaříková, R., Hůlková, I., & Schmied, J. (2020). Theme choices in Czech university students' English-medium Master's theses. Lingua, 243, 1–17. https://dx.doi.org/10.1016/j.lingua.2020.102892

Hawes, T. (2015). Thematic progression in the writing of students and professionals. Ampersand, 2, 93–100. https://doi.org/10.1016/j.amper.2015.06.002

Ho. L.V. (2011). Non-native argumentative writing by Vietnamese learners of English: A contrastive study. [Unpublished doctoral dissertation thesis]. Georgetown University. Retrieved from https://repository.library.georgetown.edu/bitstream/handle/10822/553147/hoVu.pdf

"Y'all ain't the champ of anything". Webcare goals across industries

Ursula Lutzky

Vienna University of Economics and Business

As business communication largely takes place in the digital sphere today, businesses regularly interact with their stakeholders online and thus engage in the act of webcare (van Noort and Willemsen 2012). They do so to achieve different organisational goals, including reputation and relationship management, customer care, and marketing (Van Noort et al. 2015). Previous research has explored webcare interactions for the specific style of language used (Fuoli et al. 2021; Liebrecht et al. 2021), the use of speech acts, such as apologies and complaints (Page 2014; Ruytenbeek et al. 2023), or the effect of these interactions on bystanders (Weitzl and Hutzinger 2017).

While webcare research has already yielded important insights into digital business communication, industry-specific differences in online interactions with stakeholders have not been studied extensively to date. This study explores the use of webcare across three industries: airlines, food and beverage and streaming services. In a large-scale data-driven corpus analysis, it focuses on how US companies in these industries use linguistic and communicative strategies when interacting with their stakeholders on Twitter (now X). The analysis is based on the US Corporate Twitter Corpus (UCTC) which contains 4.4m English tweets posted between September 2021 and February 2023.

Initial results show that the three industries differ in their use of webcare. While airlines mainly use webcare for customer care purposes, for the food and beverage industry and streaming services, the webcare goals of marketing and relationship management are foregrounded. By combining keyword and cluster analyses, this study explores the different uses of webcare in the UCTC to arrive at more fine-grained insights into industry-specific webcare practices and needs. It aims to show how corpus linguistic methods can be used to compare webcare communication across industries and to contribute new understanding about the different functions of webcare serving different organisational goals.

List of references

Fuoli, M., I. Clarke, V. Wiegand, H. Ziezold and M. Mahlberg (2020), 'Responding Effectively to Customer Feedback on Twitter: A Mixed Methods Study of Webcare Styles', Applied Linguistics, https://doi.org/10.1093/applin/amaa046.

Liebrecht C., C. Tsaousi and C. van Hooijdonk (2021), 'Linguistic Elements of Conversational Human Voice in Online Brand Communication: Manipulations and Perceptions, Journal of Business Research, 132: 124-135. https://doi.org/10.1016/j.jbusres.2021.03.050.

Page, R. (2014), 'Saying 'Sorry': Corporate Apologies Posted on Twitter', Journal of Pragmatics, 62: 30-45.

Ruytenbeek, N., S. Decock and I. Depraetere (2023), 'Experiments into the Influence of Linguistic (In)directness on Perceived Face-threat in Twitter Complaints', Journal of Politeness Research, 19 (1): 59–86. https://doi.org/10.1515/pr-2019-0042

Van Noort, G. and L. M. Willemsen (2012), 'Online Damage Control: The Effects of Proactive versus Reactive Webcare Interventions in Consumer-generated and Brand-generated Platforms' Journal of Interactive Marketing, 26 (3): 131-140.

Van Noort, G., L. M. Willemsen, P. Kerkhof and J. Verhoeven (2015), 'Webcare as an Integrative Tool for Customer Care, Reputation Management, and Online Marketing: A Literature Review', in P. J. Kitchen and E. Uzunoğlu (eds), Integrated Communications in the Postmodern Era, 77-99, London: Palgrave Macmillan.

Weitzl, W. and C. Hutzinger (2017), 'The Effects of Marketer- and Advocate-Initiated Online Service Recovery Responses on Silent Bystanders', Journal of Business Research, 80: 164-175.

AI-empowered TPACK for corpus technology and teacher development through online collaboration

Qing Ma[1], Xiaojuan Ma[2]

[1]The Education University of Hong Kong; [2]The Hong Kong University of Science and Technology

Teacher professional development (TPD) is crucial for teacher growth and societal advancement. The rise of AI and GenAI necessitates enhanced teacher TPACK for effective technology integration, including corpus technology. Traditional TPD methods often fall short, so we propose online collaborative learning to empower language teachers in AI-empowered TPACK for corpus technology.

We established an online Commuity of Inquiry (CoI) comprising more than 150 English teachers from more than 30 countries/regions. We explored how the CoI approach supports teachers in developing AI-empowered TPACK for corpus technology. Within this CoI, we fostered teaching presence through online workshops and a resource website to help teachers develop their corpus literacy, corpus-based language pedagogy (CBLP), and ChatGPT literacy. We encouraged teachers to interact with each other on a discussion forum to foster social presence, while cognitive presence was cultivated through members' critical inquiry and collaborative construction of AI-empowered TPACK. Teacher learning in corpus literacy, CBLP, ChatGPT literacy, and overall TPD outcomes was measured via pre- and post-study surveys over three months. Relationships between these constructs were analyzed using PLS-SEM.

104 participants completed both pre- and post-surveys. A one-way repeated measures MANOVA and the follow-up Wilcoxon signed-rank test showed significant differences between pre- and post-surveys and demonstrated teachers' improved learning outcomes in corpus literacy, ChatGPT literacy, CBLP, and TPD. The PLS-SEM results demonstrated that the teachers' corpus literacy and ChatGPT literacy significantly predicts their CBLP and TPD. Comparison between pre- and post-study surveys further revealed that, through the online CoI, participants' CBLP has become an essential predictive factor for TPD. This study demonstrates the effectiveness of online CoI for fostering collaborative TPD, empowering language teachers to co-construct AI-enhanced TPACK for corpus technology. Our research advances understanding of online CoI in developing language TPD and offers a new model for continuous teacher learning and technology integration.

Navigating gender neutrality and resistance in the Italian translation of "Pageboy": A CL analysis

Stefania M. Maci

University of Bergamo

Elliot Page's autobiography, *Pageboy*, poses a unique challenge for translators due to the author's explicit request to avoid grammatical gender in Italian when referring to their identity prior to their coming out. This requirement highlights the inherent difficulty of conveying non-binary identities in a highly gendered language like Italian, where grammatical structures often enforce binary distinctions. The linguistic tension between the author's intent and the limitations of Italian grammar serves as a focal point for exploring broader issues of gender and language (Musolff & Zinken, 2020).

To investigate these challenges, this study adopts a mixed-methods approach, combining Corpus Linguistics (CL) with sentiment analysis. The CL approach is employed to identify linguistic patterns and translation strategies used in the Italian version of *Pageboy*, focusing on adaptations for gender neutrality (McEnery & Hardie, 2021). Sentiment analysis is applied to evaluate the emotional tone and resonance of the translated text, particularly in passages where non-binary identities are discussed (Hardmeier, 2020). This dual methodology provides a nuanced understanding of how language choices impact both the accuracy of representation and the emotional engagement of the target audience.

By situating these findings within a sociolinguistic framework, the study explores the broader implications of translation practices for non-binary representation in gendered languages (Spolsky & Hult, 2020). It also interrogates the role of translators as active agents in shaping cultural narratives. Far from being neutral mediators, translators play a critical part in resisting normative linguistic frameworks and fostering inclusivity (Baker, 2021; Federici, 2020). The case study situates the Italian translation of *Pageboy* within the larger discourse on gender, language, and power, illustrating how translation practices contribute to reshaping cultural and ideological narratives. Ultimately, this paper underscores the transformative potential of translation, supported by CL and sentiment analysis, in promoting equitable representation and challenging systemic oppression (Flotow & Farahzad, 2021; Wiegand & Matlock, 2022).

List of references

Baker, M. (2021). Translation and Conflict: A Narrative Account. Routledge.
Federici, F. M. (2020). Translators, Interpreters, and Cultural Negotiators: Mediating and Communicating Power from the Middle Ages to the Modern Era. Bloomsbury Academic.
Flotow, L. von, & Farahzad, F. (2021). Translating Women: Different Voices and New Horizons. Routledge.
Hardmeier, S. (2020). "Integrating Sentiment Analysis into Translation Studies: Methodological Perspectives and Case Studies." Translation Studies Quarterly, 18(2), 135–152.
McEnery, T., & Hardie, A. (2021). Corpus Linguistics: Method, Theory, and Practice. Cambridge University Press.
Musolff, A., & Zinken, J. (2020). "Gender and Power in Translation: Linguistic and Ideological Challenges." Journal of Sociolinguistics, 24(3), 341–360.
Spolsky, B., & Hult, F. M. (2020). The Handbook of Educational Linguistics. Wiley-Blackwell.
Wiegand, F., & Matlock, T. (2022). "Emotion, Ideology, and Representation in Gendered Translations." Language and Gender Studies, 16(4), 415–432.

Ethnic disproportionality in criminal prosecution case files

Nicci MacLeod[1], Robbie Love[1], Annina Heini[2], Joyce Lim[1], Ralph Morton[1], Márton Petykó[1]

[1]Aston University; [2]University of Melbourne

The Crown Prosecution Service (CPS) prosecutes criminal cases that have been investigated by the police and other investigative organisations in England and Wales. Previous research (CPS, 2023) has revealed disproportionality in the outcomes of legal decision-making, with white British suspects less likely to be charged than suspects of other ethnic backgrounds for a comparable offence.

The CPS commissioned us to compare linguistic representations of suspects from these ethnicity groups in their case files. Using AntConc (Anthony, 2022), we analysed a 1.3-million-word corpus of files relating to 400 cases, balanced for suspect ethnicity (mixed ethnicity/white British), offence type (burglary/drugs/violence), charging status (charged/no further action), and author (police referral/CPS charging decision). We applied keyness analysis (Egbert & Biber, 2019) supplemented by qualitative analysis of transitivity (Halliday & Matthiessen, 2014) for over 3,000 concordance lines of the noun *suspect*.

Our findings reveal patterns of difference in the representation of suspects across ethnicity groups and charge status. For example, reports involving mixed ethnicity suspects contain more references to the violent or serious nature of the offence and suspects' character (cf. Cushion et al., 2007). In cases that resulted in a charge, the reports describe mixed ethnicity suspects as being more physically involved in offences, while participating less often in communicative acts, than white British suspects. Conversely, mixed ethnicity suspects who were *not* ultimately charged are represented as being involved in communicative acts more often than white British suspects.

This study demonstrates that prosecution case files are sites where an imbalance can be observed in the linguistic representation of the two selected ethnic groups. Where differences occur, they frame mixed ethnicity suspects with a higher degree of negativity and culpability than white British suspects. Such differences may implicitly perpetuate unconscious biases which can lead to tangible inequalities in terms of outcomes for suspects.

List of references

Anthony, L. (2022) AntConc (Version 4.2.0) [software]. Tokyo, Japan: Waseda University. Available from: https://www.laurenceanthony.net/software

Crown Prosecution Service. (2023) CPS action to understand disproportionality in charging decisions. https://www.cps.gov.uk/cps/news/cps-action-understand-disproportionality-charging-decisions

Cushion, J., Moore, K., & Jewell, J. (2011) Media representations of black young men and boys: Report of the REACH media monitoring project. https://orca.cardiff.ac.uk/id/eprint/28559/1/2113275.pdf

Egbert, J., & Biber, D. (2019) Incorporating text dispersion into keyword analyses. Corpora, 14(1), 77-104. https://doi.org/10.3366/cor.2019.0162.

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2004) An Introduction to Functional Grammar. Routledge.

I don't know: A corpus-based analysis of targeted lexical bundles and their role in an online anxiety support group

Corrie MacMillan[1,2], Sinéad Jackson[3], Alvin Cheng-Hsien Chen[2]

[1]University of Taipei; [2]National Taiwan Normal University; [3]University College London

As a source of authentic language data, the computer mediated communication of online support groups enables researchers to analyze the experiences of people living with mental health conditions. These groups facilitate users in motivating and encouraging each other, sharing experiences, and fostering a sense of belonging (Hwang et al., 2010; Prestin & Chou, 2014). This paper focuses on how members of r/Anxiety, an online anxiety support group with more than 738,000 anonymous members, convey their experiences of anxiety through language. This forum was selected for its size and its purpose as a place for "discussion and support for sufferers and loved ones with anxiety conditions" (Reddit, n.d.).

Data was collected from r/Anxiety using the host website Reddit's official API, resulting in a specialized corpus of 3,910 main posts containing 782,316 tokens. The corpus was analyzed using Sketch Engine's N-gram function. The most frequent 4-gram *I don't know* frequently combined with other elements to form longer 5-gram and 6-gram sequences such as *I don't know why, I don't know if, I don't know how to,* and *I don't know what to.* This paper presents a corpus-based discourse analysis of *I don't know* with a focus on these extended strings.

This research addresses the pragmatic functions of the lexical bundles and how they reflect the anxiety-related experiences of members initiating discussions. As in previous research (Bates, 2021; Collins & Baker, 2023; Pichler & Hesson, 2016), *I don't know* was used to express a lower epistemic status, as well as a prepositioned epistemic hedge (Weatherall, 2011) inviting responses. Results indicate that *I don't know* bundles helped users articulate their experiences of heightened anxiety when sharing, coping, and reaching out in initial main posts. The detailed analysis of extended *I don't know* bundles contributes an in-depth examination of their nuanced functions in anxiety discourse.

List of references

Bates, C. F. (2021). Mitigation in discourse: Social, cognitive and affective motivations when exchanging advice. Journal of Pragmatics, 173, 119-134.

Collins, L., & Baker, P. (2023). Language, discourse and anxiety. Cambridge University Press.

Hwang, K. O., Ottenbacher, A. J., Green, A. P., Cannon-Diehl, M. R., Richardson, O., Bernstam, E. V., & Thomas, E. J. (2010). Social support in an Internet weight loss community. International Journal of Medical Informatics, 79(1), 5-13. https://doi.org/10.1016/j.ijmedinf.2009.10.003

Pichler, H., & Hesson, A. (2016). Discourse-pragmatic variation across situations, varieties, ages: I don't know in sociolinguistic and medical interviews. Language & Communication, 49, 1-18.

Prestin, A., & Chou, W. Y. S. (2014). Web 2.0 and the changing health communication environment. In The Routledge handbook of language and health communication (pp. 184-197). Routledge.

Reddit. (n.d.). r/Anxiety. Retrieved December 21, 2024, from https://www.reddit.com/r/Anxiety/

Weatherall, A. (2011). I don't know as a prepositioned epistemic hedge. Research on Language & Social Interaction, 44(4), 317-337. https://doi.org/10.1080/08351813.2011.619310

Passive voice modal constructions in LModE medical writing

Martti Makinen[1], **Turo Hiltunen**[2]

[1]Hanken School of Economics; [2]University of Helsinki

Modality is an inherent element in all scientific discourse, informing the audience of the author's level of certainty, and of the levels of normativity in the discipline at hand. Different linguistic means are used to modify the information, conveying signals about the trustworthiness and plausibility of propositions (epistemic modality), or on the necessity or permissibility of ideas, actions, and events (deontic modality; Palmer 1990: 5-8, Marín-Arrese, 2009: 30, 34). Our immediate interest is the use of modal auxiliaries for these purposes.

Modality plays a key role in medical argumentation, both in historical and present-day English texts (Vihla 1999, Taavitsainen 2001). However, modality in LModE medicine has received little attention beyond our recent studies (Hiltunen and Mäkinen 2024; Mäkinen and Hiltunen 2024). These studies investigated modality in *Late Modern English Medical Writing* corpus (LMEMT, Taavitsainen and Hiltunen 2019), with the focus on active voice constructions, showing that also in medical register, modality is the aggregate effect of modal auxiliaries and lexical verbs, necessitating the phraseological approach.

The current study focuses on passive voice constructions in LMEMT. The incorporation of passive constructions complements the analysis of collocational and functional behaviour of modals in scientific argumentation. Also in passive constructions modals are associated with identifiable textual functions, e.g. MAY + KNOW in 1), relating symptoms leading to a diagnosis:

1) This fever **may** be easily **known** from the constitution of the sick person [...] (1785, Robinson: *Every Patient His Own Doctor* [general treatises])

Our study presents a colligational analysis of modals in passive clauses, focusing on their phraseological patterning with main verbs and association with specific discourse functions. We use two metrics, attraction and reliance (Schmid 2000; Hiltunen 2021) to identify MODAL + VERB pairings distinct in terms of frequencies of use and relevant to describing discourse meanings in the medical register.

List of references

Coates, Jennifer 1995. The expression of root and epistemic possibility in English. In Bybee, Joan L. and Suzanne Fleischman (eds.) Modality in Grammar and Discourse. Amsterdam: John Benjamins, 55–66. https://doi.org/10.1075/tsl.32.04coa.

Depraetere, Ilse, Bert Cappelle, & Martin Hilpert 2023. Introduction. In Depraetere, Ilse et al. (eds.) Models of Modals. From Pragmatics and Corpus Linguistics to Machine Learning. Berlin, Boston: De Gruyter Mouton, 1–13. https://doi.org/10.1515/9783110734157-001.

Hiltunen, Turo 2021. Intensification in eighteenth century medical writing. Journal of English Linguistics 49(1): 90–113. https://doi.org/10.1177/0075424220982649.

Turo Hiltunen and Martti Mäkinen. 2024. "We could scarce distinguish one from another." Towards a phraseological perspective on modal auxiliaries in three categories of Late Modern English medical writing. In Räikkönen, Jenni, Carla Suhr, Minna Palander-Collin, Arja Nurmi, Minna Nevala, Turo Hiltunen (eds.), Multilingualism and Language Variation in English across Genres and Registers: A Festschrift in Honour of Päivi Pahta. Helsinki: Modern Language Society of Helsinki, 189-218. https://doi.org/10.51814/ufy.1041.c1459.

Huddleston, Rodney 1976. Some theoretical issues in the description of the English verb. Lingua 40: 331–383. https://api.semanticscholar.org/CorpusID:170479872.

LMEMT = Taavitsainen, Irma and Turo Hiltunen (eds.). 2019. Late Modern English medical texts: Writing medicine in the eighteenth century. Amsterdam: John Benjamins.

Mäkinen, Martti and Turo Hiltunen. 2024. Modal auxiliaries in four medical text categories in the LModE period. Paper read at Eighth International Conference on Late Modern English, Salamanca, Spain, Oct 2-4, 2024.

Marín-Arrese, Juana Isabel. 2009: Effective vs. Epistemic Stance, and Subjectivity/Intersubjectivity in Political Discourse. A Case Study. In Tsangalidis, A. and Facchinetti, R. (eds.). Studies on English modality. In honour of Frank R. Palmer. Berlin: Peter Lang, 23-52.

Palmer, Frank Robert. 1990. Modality and the English Modals. 2nd ed. London and New York: Longman.
Schmid, Hans-Jörg 2000. English Abstract Nouns as Conceptual Shells: From Corpus to Cognition. Berlin: Mouton de Gruyter.
Sinclair, John 2008. The phrase, the whole phrase, and nothing but the phrase. In Meunier, Fanny and Sylviane Granger (eds.) Phraseology: An Interdisciplinary Perspective. Amsterdam: John Benjamins, 407–410.
Sweetser, Eve 1990. From Etymology to Pragmatics: Metaphorical and Cultural Aspects of Semantic Structure (Cambridge Studies in Linguistics). Cambridge: Cambridge University Press.
Taavitsainen, Irma. 2001. Evidentiality and scientific thought-styles: English medical writing in Late Middle English and Early Modern English. In Maurizio Gotti and Marina Dossena (eds.), Modality in specialized texts: Selected papers of the 1st CERLIS Conference. Bern: Peter Lang, 21–52.
Vihla, Minna.1999. Medical Writing: Modality in Focus. Amsterdam: Rodopi.
Warchał, Krystyna 2015. Certainty and Doubt in Academic Discourse: Epistemic Modality Markers in English and Polish Linguistics Articles. Katowice: Wydawnictwo Uniwersytetu Śląskiego.

From prototype to paradox: The evolution of the lone-wolf terrorist in newspaper discourses (WITHDRAWN)

Daniel Malone

Edge Hill University

Recent acts of extremist violence in Magdeburg (Germany), New Orleans (USA), and Southport (UK) have once again raised concerns about self-radicalisation and lone-wolf (i.e., lone-actor) terrorism. In a statement regarding the Southport event, UK Prime Minister Keir Starmer stated that "terrorism has changed" and that Britain faces a "new and dangerous threat" from "loners, misfits, young men in their bedroom" (Rhoden-Paul, 2025). This paper explores how the lone-wolf terrorist's (LWT) characteristic 'aloneness' is represented in the UK press and whether it is connected to other actors, networks, and organisations.

The dataset for this paper was the Lone Wolf Corpus (Malone, 2020), a topic-specific 8.5-million-word corpus of UK newspaper articles published between 2000 and 2019. Adopting a semasiological approach, it employs prototype-structured categorisation (e.g., Lakoff, 1987) to identify and analyse connection types through their defining attributes. This componential analysis reveals distinct profiles of connection types, with the Prototypical LWT depicted as ideologically self-driven and operationally independent. Treating these connection types as discursive variables enabled statistical analyses of their frequency to identify shifting diachronic patterns.

Findings indicated that the LWT underwent a discursive reconstruction marked by institutionalisation and depersonalisation. Early representations more frequently aligned with the Prototypical LWT, while later portrayals increasingly emphasised links to broader networks, often depicting the LWT as faceless, anonymous agents mobilised by external directives from Islamist organisations. These shifts reflected evolving attitudes concerning who or what constituted the "terrorist threat" of the time and illustrated how the LWT was discursively shaped and reshaped depending on sociopolitical and security priorities.

By integrating prototypical categorisation into corpus-assisted discourse analysis, this study provides a novel methodological framework for investigating dynamic discourses. This approach offers insights for researchers examining discourse with corpora, particularly from a diachronic perspective.

List of references

Lakoff, G. (1987). Women, fire, and dangerous things: What categories reveal about the mind. Chicago: University of Chicago Press.

Malone, D. (2020) Developing a complex query to build a specialised corpus: Reducing the issue of polysemous query terms. Paper presented at Corpora and Discourse International Conference 2020, University of Sussex, UK.

Rhoden-Paul, A. (2025, January 22). 'Terrorism has changed', says PM on Southport attacks. BBC News. https://www.bbc.co.uk/news/articles/cvg9p39kez7o

British news media representations of the 2022 and 2024 Mpox outbreaks: A corpus-assisted study

Beth Malory[1], Marthe Le Prevost[1], Emily Jay Nicholls[1], Davide Bilardi[2], Shema Tariq[1]

[1]University College London; [2]Paediatric European Network for Treatment of AIDS (PENTA)

May 2022 saw the first cluster of UK cases of non-travel associated mpox, a virus previously endemic only in western and central Africa. Since then, over 100,000 people across 100 countries have been diagnosed with mpox, with the World Health Organization (WHO) declaring it a Public Health Emergency of International Concern (PHEIC) in July 2022. This evolving outbreak, coming swiftly on the heels of COVID-19 pandemic, attracted considerable global media attention. By late 2022, case numbers were falling and mpox was declared no longer a PHEIC. However, a new variant (clade) of mpox, identified in 2023, has resulted in another upsurge of cases in Africa, with the WHO once again declaring mpox a PHEIC in 2024.

Using diachronic frequency analysis and Corpus-Assisted Discourse Studies, we analysed UK news media sub-corpora of the English Trends corpus on SketchEngine (*English Trends*, 2024). We compared how British news reportage represented mpox during the 2022 and 2024 outbreaks, focusing on representation of minoritised groups, employing a mixed-methods approach.

Representation in the two outbreaks was notably distinct. Frequency analysis shows that mpox attracted significantly more coverage in 2022; 2022 hits for 'monkeypox' (then accepted nomenclature) account for almost ten times the 2024 raw frequency of 'mpox'. Collocational analysis demonstrates mpox's construction in 2022 as both an imminent threat in Britain and one associated with sexual promiscuity and men who have sex with men. In 2024, mpox was represented less as a threat to UK residents, and instead as an issue primarily affecting the Global South. Transitivity analysis of concordances shows that, whilst distancing mechanisms were utilised differently and to different degrees during each outbreak, discursive othering remained a consistent strategy. This is likely to play an important role in the stigmatisation of mpox, ultimately hampering public health efforts to control outbreaks.

List of references

English Trends. (2024). SketchEngine. https://www.sketchengine.eu/

The German "gehören"-passive – or: How to research non-canonical constructions in language corpora

Claudia Mattes

University of Vienna

The German sentence structure – among other peculiarities of the language – inspired Mark Twain to rant extensively in his essay about “The Awful German Language”. It also poses a challenge when researching verb complexes at the intersection of morphology and syntax.

A less prominent example of this is the so-called *gehören*-passive. Some of its examples show the parts of the construction in sequence, …

*Das Gesetz **gehört geändert**.* ('The law **should be changed**.')

…, while most include objects and attributes between the finite and infinite parts of the verb.

*Es **gehören** sicher ein paar Spieler **ausgetauscht**.* ('Surely a few players **should be replaced**.')

Semantically, the construction of *gehören* ‘belong to’ and a past participle expresses passivity and deontic modality at the same time, with the meaning that something ought to be done while the agent is not in focus due to deagentivation. Its grammaticalisation has only been studied at by a few researchers so far (Szatmári 2002; Stathi 2010; Lasch 2018; Mattes 2024).

Most available NLP tools are based on large corpora that often represent certain standard(ised) varieties and usually aren’t well suited for the study of non-canonical constructions, even for larger languages such as German (see f.e. Blaschke et al. 2024), with its variation across the whole spectrum between dialect and standard (as it is common for pluricentric languages).

This contribution addresses the challenge of finding (syntactically and semantically) complex constructions that are not annotated in larger language corpora, by analyzing the occurrence of the *gehören*-passive within the Austrian Media Corpus (amc), the largest resource of Austrian Standard German (ASG) consisting of most (print) media in Austria over the last 30 years.

Early results show distinct differences between the *gehören*-passive and other constructions with *gehören* as well as the need for further research into different corpora.

List of references

Austrian Media Corpus (amc). Retrieved from https://amc.acdh.oeaw.ac.at/ [15.1.2025]

Blaschke, Verena, Barbara Kovačić, Siyao Peng, Hinrich Schütze & Barbara Plank. 2024. MaiBaam: A Multi-Dialectal Bavarian Universal Dependency Treebank. Retrieved from https://arxiv.org/abs/2403.10293v1 [15.1.2025]

Dorn, Amelie, Jan Höll, Theresa Ziegler, Wolfgang Koppensteiner & Hannes Pirker (2023). Die österreichische Presselandschaft digital: Das Austrian Media Corpus (amc) und sein Potential für die Linguistik. In Marc Kupietz & Thomas Schmidt (ed.), Neue Entwicklungen in der Korpuslandschaft der Germanistik. Beiträge zur IDS-Methodenmesse 2022. (= Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache (CLIP) 11), 43–55. Tübingen: Narr.

Lasch, Alexander. 2018. „Diese gehören kalt zu geben.“ Die Konstruktion „gehören“ mit Qualitativ. In Sprachwissenschaft 43 (2), 159–185.

Mattes, Claudia (2024): Das gehören-Passiv in der österreichischen Standard(schrift)sprache. Eine Analyse im Austrian Media Corpus (amc). Masterarbeit Universität Wien. Philologisch-Kulturwissenschaftliche Fakultät.

Ransmayr, Jutta, Karlheinz Mörth & Matej Ďurčo. 2017. AMC (Austrian Media Corpus) – Korpusbasierte Forschungen zum österreichischen Deutsch. In Claudia Resch & Wolfgang U. Dressler (ed.), Digitale Methoden der Korpusforschung in Österreich (= Veröffentlichungen zur Linguistik und Kommunikationsforschung Nr. 30), 27–38. Wien: Verlag der Österreichischen Akademie der Wissenschaften.

Stathi, Katerina. 2010. Is German gehören an auxiliary? The grammaticalization of the construction gehören + participle II. In Katerina Stathi, Elke Gehweiler & Ekkehard König (ed.), Grammaticalization: current views and issues. Studies in language companion series (SLCS), 323–342. Amsterdam & Philadelphia: Benjamins Pub. Co.

Szatmári, Petra. 2002. Das gehört nicht vom Tisch gewischt... Überlegungen zu einem modalen Passiv und dessen Einordnung ins Passiv-Feld. In Jezikoslovlje 3(1–2), 171–192.
Twain, Mark. 1880. The Awful German Language. Published by U.S. Embassy, Berlin, Germany in 2010. Retrieved from https://www.daad.org/files/2022/09/Mark_Twain-Broschuere.pdf [15.1.2025]

'I want to protect my loved ones': A corpus linguistic analysis of vaccine perceptions to inform a pan-London immunisation campaign

Emma McClaughlin[1], Svenja Adolphs[1], Sara Vilar-Lluch[2], Ysabella Hawkings[3]

[1]University of Nottingham; [2]Cardiff University; [3]Association of Directors of Public Health (ADPH) for London

Increasing immunisation uptake is an urgent global health priority according to the World Health Organisation [1]. Vaccine uptake in London, a region with a highly mobile population and social inequalities, is lower than the England average leaving millions at risk. Responding to increasing outbreaks of vaccine-preventable illnesses including measles, mumps, and rubella, the UK Health Security Agency's new pan-London campaign 'Why we get vaccinated' has been designed collaboratively with corpus linguists to encourage open conversations about vaccination in the region. This paper reports a corpus linguistic analysis of open-text responses to a survey of vaccination attitudes conducted in August 2024 with 2,053 London adults. Keywords and key multiword terms were extracted from responses to questions on motivations for seeking the flu vaccine and for deciding to give children the MMR vaccine, followed by a qualitative examination of concordance lines. Findings revealed that past experiences of flu motivated vaccine uptake, respondents recognised community benefits of herd immunity, and strong support for vaccination relates to personal responsibility and trust in vaccine safety. Language associated with vaccine support includes 'precaution', 'protect', 'risk of severe illness'. Barriers to uptake include perceived lack of need and costs associated with vaccinating. We also found 'othering' of non-vaccinators and a general, rather than specific, understanding of risks in vaccine positive respondents. We discuss whether leveraging certain language choices can reinforce prosocial benefits of vaccination [2], as well as the misconceptions about immunity and vulnerability that need to be addressed. Vaccine-positive individuals often 'believe' and 'hope' that they will be protected, rather than 'know' it, so this group can also benefit from targeted messaging to prevent slippage of vaccine positivity. The findings will support asset distribution and evaluation and will inform the iterative design of future assets for the campaign, which is set to run across London until 2027.

List of references

[1] World Health Organisation (2020) "Immunization Agenda 2030." https://www.who.int/immunization/immunization_agenda_2030/en/

[2] Vilar-Lluch, S., McClaughlin, E., Knight, D., Adolphs, S., & Nichele, E. (2023). The language of vaccination campaigns during COVID-19. Medical Humanities, 49(3), 487-496.

Online discourses of 'Gender Critical' feminists: A keyword co-occurrence analysis

Mark McGlashan[1], **Isobelle Clarke**[2], **Tony Berber-Sardinha**[3], **Claudia Delfino**[3], **Mirella Whiteman**[3]

[1]University of Liverpool; [2]Lancaster University; [3]Pontifical Catholic University of São Paulo

TW: gender, violence, hate speech

Trans people - people whose assigned biological sex does not conform to their gender identity - and trans women especially, have been subject to "sustained antagonism from sections of feminism" (Hines 2017). These sections of feminism, self-described as 'gender critical' but commonly referred to as Trans Exclusionary Radical Feminists (TERFs), have become central to contemporary (often polemical) public debates on gender-based discrimination and inequity. Terms like "panic" (Thomsen & Essig 2020) and "war" (Pearce, et al. 2020) have come to characterise the polarising debate around trans exclusion and, according to Pearce et al. (2020: 684), "[t]he TERF wars [...] are best understood as a series of complex discursive and ideological battles within (rather than against) feminism". In June 2020, r/GenderCritical - a major online anti-trans space - was banned from the social media site Reddit for hate speech (Tiffany, 2020) and migrated to https://ovarit.com (Ovarit referencing ovaries as a way to exclude any person not born with ovaries, i.e. trans women).

For this paper, we collected a corpus of all 504 posts (including their comments) to the Ovarit subforum /o/GenderCritical that were available at the time of collection (July 2023). This study explores the linguistic dimensions underlying the GenderCritical circle of Ovarit using Keyword Co-occurrence Analysis (KCA; Clarke et al., 2021) to examine recurring discourses. KCA is an approach that applies Multiple Correspondence Analysis (MCA) to identify groups of keywords based on their frequent co-occurrences in texts. Our findings reveal patterns of keyword co-occurrence that underpin a range of overlapping yet distinct discourses that characterise gender-critical and radical feminist discussions in this online forum, including 'biological essentialism', 'feminist resistance/protection of feminism and feminists', 'gender performance', and '(male) violence and abuse'.

List of references

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.

Hines, S. (2019) The feminist frontier: on trans and feminism. Journal of Gender Studies 28(2): 145-157. doi: 10.1080/09589236.2017.1411791

Kilgariff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. and Suchomel, V. (2014) The Sketch Engine: Ten years on. Lexicography 1: 7—36.

Pearce, R., Erikainen, S. and Vincent, B. (2020) TERF wars: An introduction. The Sociological Review 68(4): 677–698. doi: 10.1177/0038026120934713.

Thomsen, C. & Essig, L. (2022) Lesbian, feminist, TERF: a queer attack on feminist studies, Journal of Lesbian Studies 26(1): 27-44. doi: 10.1080/10894160.2021.1950270

Tiffany, K. (2020) The secret internet of TERFs. The Atlantic. Accessed on 12th April 2022 from: https://www.theatlantic.com/technology/archive/2020/12/reddit-ovarit-the-donald/617320/

Discussing men's rights online: Comparing the language of pro- and anti-feminist online communities

Mark McGlashan[1], **Alexandra Krendel**[2]

[1]University of Liverpool; [2]University of Southampton

We examine linguistic similarities between a pro-feminist online men's community (r/MensLib) and an anti-feminist online men's community (r/MensRights) on Reddit. Current research focuses on the latter kind of community because of links to violent misogyny (e.g. the involuntary celibate community; *incels*). However, we aim to explore areas of potential common interest between these communities to explore routes to productive discussion around - and speculative de-escalation from - harmful forms of misogyny in online anti-feminist communities, thus aiming to challenge violence against women and girls (VAWG).

We created two subcorpora by collecting all comments from the 500 most recently published threads in r/MensLib and r/MensRights (collected 28/11/2024) alongside a reference corpus of 100,000 comments collected from 31,731 different subreddits, which aims to represent the language found in comments across Reddit.[1] We follow methods described in McGlashan & Krendel (2023) to compare language used in r/MensLib and r/MensRights by identifying key-key-word lists for each subcorpus before comparing them to identify shared (and distinctive) lexical items. Key-key-words were identified by first producing a keyword list for each of the 500 threads in each subcorpus using the Conservative LogRatio statistic (LRC; Evert 2021); only keywords meeting LRC $\geqslant$2 were kept. Key-key-words were those keywords found in $\geqslant$5% (25/500) of threads in each subcorpus. For the purpose of this paper, we compare the top 100 key-key-words matching these cutoffs.

r/MensRights-specific key-key-words concern **subjugation of men** (*misandry*), **(sexualised) violence** (*rape, assault, dv* [domestic violence]), **genitals** (*circumcision*), and **(reproductive) biology** (*paternity, abortion*). r/MensLib-specific key-key-words, on the other hand, concern **politics** (*alt-right, leftist*), **emotion** (*emotions, empathy*), and **marginalisation** (*loneliness, marginalized*). Both focus on **gender** (*gender, masculinity, feminity*), **sexism** (*sexism, misogyny*), **(in)equality** (*equality, oppression*), **feminism** (*feminists*), and **incels**.

[1] Reference corpus comments were extracted from a total capture of all comments and submissions to Reddit during November 2024 https://academictorrents.com/details/a1b490117808d9541ab9e3e67a3447e2f4f48f01?stats=True

List of references

Evert, S. (2022) 'Evert (2022): Measuring keyness'. Available at: https://doi.org/10.17605/OSF.IO/CY6MW.

McGlashan, M., & Krendel, A. (2023). 'Keywords of the manosphere'. International Journal of Corpus Linguistics. https://doi.org/10.1075/ijcl.22053.mcg

Investigating Discursive Justifications of Equitable Prescriptivism in the Equal Treatment Bench Book (2013-2024): A corpus-assisted discourse analysis

Jamie McKeown

The City University of Hong Kong

First published in 2013 by the Judicial College of England and Wales—a body responsible for the professional development of judges—the Equal Treatment Bench Book (ETBB) aims to guide the judiciary in effectively communicating with legal actors from diverse backgrounds. Amongst other things, it offers frameworks for appropriate language use, cultural awareness, and inclusive practices. Despite its commendable aims, the ETBB has faced significant criticism, resulting in five revised editions and two interim versions. Using a series of corpora comprising all editions published thus far (1,204,654 words in total), this study will investigate a primary criticism directed at the ETBB: lack of justification.

Justification is essential to the enactment of judicial authority. Advocates present their arguments with persuasive justifications (Johnstone, 2002), and judges are expected to clearly justify their decisions (Feteris, 1999: 6). Acts of judicial fiat often encounter significant opposition. Considering the significance of justification in upholding judicial legitimacy, it is essential to understand how the writers of the ETBB engage in this discursive practice. The present study employs a corpus-assisted discourse analytic approach to investigate the explicit attempts at justification made by the writers of the ETBB. Specifically, it will identify markers of justification (e.g. see, Hyland, 2005; Van Eemeren et al., 2007) and classify the wider concordances in which they occur according to MacCormick's (1994) typology of legal justifications.

Preliminary observations indicate that justifications related to consequences (i.e., arguments based on acceptable or unacceptable societal outcomes) occur more frequently than those focused on coherence or consistency (i.e., alignment with general legal principles or specific rules). In other words, the writers of the ETBB prioritise societal consequences over legal fit (see Posner, 1995: 11). Diachronically, explicit attempts at justification increase over time, indicating a growing sensitivity among the writers to the expectations of the wider epistemic community.

List of references

Feteris, E.T. 1999. Fundamentals of Legal Argumentation: A Survey of Theories on the Justification of Judicial Decisions. Kluwer Academic Publishers.
Hyland, K. 2005. Metadiscourse. Continuum.
Johnston, T.R. 2002. Oral Arguments and Decision Making on the United States Supreme Court. SUNY Press.
MacCormick, N. 1994. Legal Reasoning and Legal Theory. OUP.
Posner, R.A. 1995. Overcoming Law. Harvard University Press.
Van Eemeren, F.H., Houtlosser, P., & Snoeck Henkemans, A.F. 2007. Argumentative Indicators in Discourse. A Pragma-Dialectical Approach. Springer.

ParlaTalk: An automated system for continuous development of European parliamentary corpora

Ota Mikušek[1,2]

[1]Faculty of Informatics, Masaryk University; [2]Lexical Computing, Brno

This paper focuses on designing, implementing, and maintaining a system for the continuous automatic development of European parliamentary corpora called ParlaTalk. Parliamentary protocols from EU member states are a rich and continuously expanding source of multilingual textual transcriptions of spoken language that hold significant value for linguistic research, machine learning, political analysis, and journalism. However, the level of accessibility of these data varies in each parliament chamber because of inconsistencies in formats, metadata availability, and different data access methods.

To solve these problems, ParlaTalk was developed to automate the downloading and processing of parliamentary protocols into standardized corpora. ParlaTalk is a set of tools that transform data from various sources into a unified format while preserving original information. ParlaTalk gathers data from 22 EU member states, resulting in 22 continuously updated corpora. These corpora are enriched with sheared metadata, including speaker names, session dates, source URLs, and transcriber notes. The created corpora are compatible with (No)Sketch Engine and have been made publicly available through the Sketch Engine platform.

Currently, this system is running continuously for 2 years. What makes this possible are error detection functions and recovery mechanisms that react to source changes in data format or outages. Evaluation of the resulting corpora involved checks for data integrity and statistical assessments.

ParlaTalk demonstrates significant improvements in automating the collection and processing of parliamentary texts, contributing to computational linguistics and political discourse studies. Future work will focus on expanding the system to include non-EU parliamentary data, extending error-handling mechanisms, and expanding the size of metadata annotations.

PseudoBrown: An AI-generated corpus for comparative analysis of human and LLM-authored texts

Jiří Milička, Anna Marklová, Václav Cvrček

Charles University

Corpus linguists should not ignore developments in the field of large language models (LLMs), even if they are solely interested in purely human language in human-produced texts. Generated texts are becoming ubiquitous and will inevitably influence human perception and production, whether we like it or not. Currently, linguistic studies of LLM-generated texts are largely experimental, such as attempts to adapt classical psycholinguistic experiments (Weissweiler et al. 2013, among many others). Corpus linguistics should not take a back seat, since it has potential to apply quantitative and corpus-based paradigms and methodologies.

A key challenge today is the scarcity of suitable data. Ideally, researchers would need an open corpus of LLM-generated texts comparable to an established human-generated corpus, which could serve as a control dataset, a reference corpus.

Our newly created corpus, *PseudoBrown* (as the name suggests), is modeled on the Brown Corpus (Francis – Kucera 1979), with which it is directly comparable. The Brown Corpus was chosen because it contains excerpts from a wide variety of genres. Though dated, this is advantageous: while it likely appears in LLM training data, its age ensures it is purely human-generated—free from machine-generated texts, whether directly authored by LLMs or via machine translation.

The *PseudoBrown* corpus was generated using tens of models (by OpenAI, Anthropic, Meta, DeepSeek, and Alphabet) under various configurations and tagged according to the Universal Dependencies standards (Nivre et al. 2020).

The presentation will introduce the corpus in detail and demonstrate its applications. Examples include analyses using stylometry, keyword extraction, grammatical bias detection, and other classic corpus linguistic methods.

The corpus will be published as an open dataset by the time of the conference.

List of references

Francis, W. N., & Kucera, H. (1979). Brown corpus manual. Letters to the Editor, 5(2), 7.

Nivre, J., De Marneffe, M. C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., ... & Zeman, D. (2020). Universal Dependencies v2: An evergrowing multilingual treebank collection. arXiv preprint arXiv:2004.10643.

Weissweiler, L., Hofmann, V., Kantharuban, A., Cai, A., Dutt, R., Hengle, A., ... & Mortensen, D. R. (2023). Counting the Bugs in ChatGPT's Wugs: A Multilingual Investigation into the Morphological Capabilities of a Large Language Model. arXiv preprint arXiv:2310.15113.

From Synchronic to Diachronic: Evolving collocates of 'Novel' in biomedical texts

Neil Millar[1], **Maria Munteanu**[2]

[1]University of Tsukuba; [2]Lebanon High School

In scientific discourse, the use of the term *novel* has increased significantly in recent years. For instance, in 1985, 9% of abstracts describing projects funded by the U.S. National Institutes of Health (NIH) contained the term; by 2020, this figure rose to 36%. [1] This trend has been linked to the broader phenomenon of increasing promotion in science (*hype*) but also suggests underlying changes in the meaning of the adjective. While collocation analysis, a standard methodology in corpus linguistics, can indicate how aspects of a word's meaning emerge from recurring contexts, it is typically applied to synchronic data. This paper explores the application of collocation analysis to diachronic corpora to investigate changes in the meaning of *novel* in biomedical texts by analysing noun collocates over time.

Using abstracts from all NIH grant applications between 1985 and 2020 (901,717 abstracts), we extracted 320,355 sentences containing the term *novel*. Employing traditional NLP tools (SciSpacy) and a fine-tuned language model (ChatGPT), we first identified nouns directly related to *novel* and categorized their syntactic relationship as predicative (*the method is novel*) or attributive (*the novel method*). For each year, we then calculated association metrics (mutual information, log-likelihood, odds ratio, Delta-P, and t-score) to assess patterns of change over time.

Findings indicate: (1) a shift towards attributive usage, suggestive of increasing promotional use; (2) semantic broadening; (3) shifts in the semantic categories of noun collocates, from, for example, research-oriented terms (*approach, method*) to outcome-oriented terms (*findings, results*); (4) the emergence of strong collocations, suggesting the development of formulaic expressions in grant writing; and (5) the appearance of new collocations in later years. We discuss findings in relation to evolving norms of scientific communication. Additionally, we discuss challenges in applying diachronic collocation analysis to large-scale corpora and consider the advantages of using LLMs in corpus research.

List of references

[1] Millar, N., Batalo, B., & Budgell, B. (2022). Trends in the use of promotional language (hype) in abstracts of successful national institutes of health grant applications, 1985-2020. JAMA network open, 5(8), e2228676-e2228676.

The language of hate: Understanding linguistic variations with Turkish Tweet analysis

Hülya M□s□r, Jack Grieve

University of Birmingham

Problematic online behavior, including hate directed at groups or institutions and interpersonal harassment or bullying, is increasingly recognized as a precursor to hostility, violence, and societal discord. Social media platforms, while ubiquitous and convenient for content generation, are frequently exploited to propagate aggression and hatred. While computational tools for monitoring online abuse and hate speech have increased, the linguistic and stylistic patterns underpinning such language remain underexplored.

This study investigates the structural patterns driving hate speech and offensive language in two prominent Turkish Twitter corpora with diverse labeling systems: Toraman et al. (2022), with 60,310 tweets labeled as hateful, offensive, or neutral distributed over five domains, and Çöltekin (2020), containing 35,015 tweets categorized as hate speech by target (group, individual, institutional), profanity, or neutrality. We developed grammatical feature sets in Turkish for each corpus and applied Multiple Correspondence Analysis, a multidimensional analysis of short texts, where neutral tweets defined the principal dimensions, and tweets containing hate speech and offensive language were projected onto the model to assess their alignment with this structure.

Our analysis identified the first dimension as a reliable indicator of tweet length, serving as a confounding variable. Distinct differences in interactive style, action-orientedness, and outward-facing discourse characterized hate and offense subsets, with networked communication features further contributing to the fostering of collective aggression. In contrast, differences in aspectual and temporal specificity and narrativity were less pronounced. Temporal analysis revealed consistent stylistic patterns over time, highlighting the enduring influence of linguistic structures in driving online hostility and providing insights for strategies to combat online abuse. These findings showcase linguistic dimensions critical to understanding online hate and offer a foundation for improving computational tools to detect and mitigate such behavior in Turkish social media.

List of references

Çöltekin, C., (2020). A Corpus of Turkish Offensive Language on Social Media. In Proceedings of the Twelfth Language Resources and Evaluation Conference, (pp. 6174–6184), Marseille, France. European Language Resources Association. https://aclanthology.org/2020.lrec-1.758.

Toraman, Ç, Şahinuç, F., & Yilmaz, E. (2022). Large-Scale Hate Speech Detection with Cross-Domain Transfer. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, (pp. 2215–2225), Marseille, France. European Language Resources Association. https://aclanthology.org/2022.lrec-1.238.

Predicting the use of dative-experiencer constructions over nominative-experiencer constructions: A diachronic analysis of Spanish constructions of 'liking'

Andrea Celeste Mojedano Batel

Aston University

This corpus study, spanning the 13th to the 19th centuries, examines the diachronic change in argument structure in Spanish constructions of 'liking,' focusing on constructions containing the verb *gustar* 'to like.' Data come from a three-million-word corpus with 41 digitized Peninsular Spanish narrative texts.

Gustar shifted its argument structure, transitioning from a nominative-experiencer pattern to a dative-experiencer pattern. Verbs of this type diverge from the standard transitive pattern because their grammatical subject (which has the role of stimulus) does not correspond to the logical subject (which has the role of experiencer), as in *Me.OBJECT gustan los sermones.SUBJECT* 'I like preachings/sermons.'

I have previously argued that the semantico-syntactic properties of the stimuli in dative-experiencer constructions helped trigger processes of change in argument structure (Mojedano Batel, 2020; in press). This study builds upon said research by considering further morphosyntactic and semantic factors driving constructional change. New morphosyntactic factors include the verb's grammatical aspect, as dative-experiencer constructions generally express atelic states of affairs (Miglio et al., 2013; Rivas, 2016). New semantic factors include polarity, as the notion of how negation affects the argument structure of *gustar* has not been studied before, and causation. By incorporating further morphosyntactic and semantic perspectives, I provide a more comprehensive account of the factors driving constructional change.

To understand which linguistic variables predict the use of one argument structure construction or another, I use a generalized estimating equation (GEE), applied to data from each time period separately. All fixed-effects variables (i.e., *Animacy of stimulus*, *Syntactic category of stimulus*, *Grammatical Aspect*, *Causation*, and *Polarity*) proved significant in one or more of the time periods under analysis.

Findings shed light on cross-linguistic variation in argument structures and extend scholarship on the issue of word order in psychological verb constructions and on how negation affects variation in argument structure.

List of references

Miglio, V., Gries, S., Harris, M., Wheeler, E., & Santana-Paixão, R. (2013). Spanish lo(s)-le(s) Clitic alternations in psych verbs: A multifactorial corpus-based analysis. In J. Cabrelli Amaro, G. Lord, A. de Prada Pérez, & J. E. Aaron (Eds.), Selected proceedings of the 16th Hispanic linguistics symposium (pp.268–278). Somerville, MA: Cascadilla Proceedings Project.

Mojedano Batel, A. (2020). 'Liking' constructions in Spanish: Syntactic category of the stimulus and constructional change. In Drinka, Bridget (Ed.), Historical Linguistics 2017. Amsterdam: John Benjamins.

Mojedano Batel, A. (In press). Diachronic Change in Spanish 'Liking' Constructions: A Case of Analogical Extension Through a Multiplicity of Source Constructions. Journal of Historical Linguistics.

Rivas, J. (2016). Verb–object compounds with Spanish dar 'give': an emergent gustar 'like'-type construction, WORD, 62(1), 1-21, DOI: 10.1080/00437956.2016.1141940

The EXEMPRAES Corpus: A genre-based tool for teaching and investigating English vs. Spanish academic writing

Ana I. Moreno

UNIVERSIDAD DE LEON

Corpus tools have long been recognized as useful for teaching academic writing, particularly in multilingual and cross-disciplinary contexts. Building on this, I argue that such tools can also foster greater awareness of cross-linguistic and cross-cultural variation, promoting deeper learning. Yet, conventional corpus tools have key limitations: they typically focus on English, rarely link linguistic features to rhetorical purposes, and offer little support to learners unfamiliar with the target language—who cannot search for what they do not yet know. Past attempts to annotate texts for rhetorical structures often rely on the sentence as the unit of analysis, a method that oversimplifies academic discourse. Since sentences often serve multiple rhetorical functions, sentence-based annotations fall short—especially for cross-cultural comparisons, where information packaging differs widely. The ***EXEMPRAES Corpus*** (Exemplary Empirical Research Articles in English and Spanish), developed under the **ENEIDA Project**, addresses these challenges with a functional, cross-linguistic approach to annotation. It uses Moreno and Swales' (2018) contrastive move-analysis framework and adopts the **meaningful proposition** as its unit of analysis. The *EXEMPRAES Corpus* distinguishes **moves-steps proper**, **announcements**, and **elaborations**, enabling finer-grained, pedagogically relevant analysis. This structure supports learners in identifying communicative functions in one language and locating equivalent patterns in the other—starting with what they know. The new **Segment Bank** offers a searchable repository of annotated examples by function, further supporting this process. Also new, the **Key Segment in Context (KSIC)** function displays annotated segments within their rhetorical, disciplinary, and textual environment. Complementary tools (KWIC and Word List) allow for exploration of phraseology and frequency across the same rhetorical function. While still modest in size, the *EXEMPRAES Corpus* is a **proof of concept**—demonstrating how functionally annotated, contrastive corpora can support both research and the teaching of academic writing across languages and cultures.

(In dedication to **John Swales**, whose insights inspired the *EXEMPRAES Corpus*)

List of references

Chen, Meilin and Flowerdew, John. (2018). Introducing data-driven learning to PhD students for research writing purposes: A territory-wide project in Hong Kong. English for Specific Purposes, 50: 97–112.

Flowerdew, L. (2022). Using corpora for writing instruction. In A. O'Keeffe & M. McCarthy (Eds.), The Routledge Handbook of Corpus Linguistics (2nd ed., pp. 443–450). Routledge. https://doi.org/10.4324/9780367076399-31.

Moreno, A. I. 2021a. Selling research in RA discussion sections through English and Spanish: An intercultural rhetoric approach. English for Specific Purposes, 63, 1-17. https://doi.org/10.1016/j.esp.2021.02.002.

Moreno, A. I., & Swales, J. M. (2018). Strengthening move analysis methodology towards bridging the function-form gap. English for Specific Purposes, 50, 40-63.

Interdisciplinarity in digital humanities journals: A corpus-based study of status markers

Natalia Muguiro

National University of La Pampa.

The importance of interdisciplinarity in prompting innovation in both research and pedagogy has been widely acknowledged (Frodeman et al., 2017). In comparison with previous work on disciplinary differences, little research exists on interdisciplinary discourse (see Choi & Richards, 2017; Muguiro, 2020, 2022; Oakey & Russell, 2014; Thompson et al., 2017; Thompson & Hunston, 2020). This paper thus addresses the call for corpus-based studies on interdisciplinary academic writing. Defined as an interdisciplinary field per se (Davidson & Savonick, 2017) and due to its growing relevance as an area of research (Adolphs & Knight, 2020), Digital Humanities constitutes a valuable case to study. For this, a corpus of 400 Research Articles from four different Digital Humanities journals (Spinaci et al., 2022) was built to explore epistemic status (Hunston, 2000, 2011). As status given to propositions indicates the ideology surrounding the research process, evaluations of status help to describe disciplines and their assumptions as well as relations between disciplines (Thompson & Hunston, 2020). Adopting a top-down approach, both quantitative and qualitative methods were applied. For comparative purposes, the frequency of selected status markers was calculated across the four journals followed by a more fine-grained qualitative analysis of the most salient markers. The results were interpreted in the light of epistemological frameworks. Status markers served to describe different typologies of interdisciplinarity by showing traces of 'methodological', 'theoretical', 'instrumental', and 'critical' types (Klein, 2015, 2017). Furthermore, different modes of engagement (Svensson, 2010) between the disciplines involved were identified. Perhaps more than any area, Digital Humanities has succeeded in linking disciplines that are radically disparate in focus and methodology (Davidson & Savonick, 2017). The study of status markers constitutes one possible way to show how these disciplines interact when forming the interdisciplinary field of Digital Humanities and how language reflects (and is reflected by) these interactions.

List of references

Adolphs, S., & Knight, D. (2020). English language and the digital humanities. In S. Adolphs, & D. Knight (Eds.), The Routledge Handbook of English Language and Digital Humanities (pp.1-4). London: Routledge.

Choi, S., & Richards, K. (2017). Interdisciplinary discourse: Communicating across disciplines. London: Palgrave.

Davidson, C., & Savonick, D. (2017). Digital Humanities: The role of Interdisciplinary humanities in the information age. In R. Frodeman, J. Klein, & R. Pacheco (Eds.), The Oxford Handbook of Interdisciplinarity (2nd ed., pp. 159-172). Oxford: Oxford University Press.

Frodeman, R., Klein, J., & Pacheco, R. (2017). The Oxford Handbook of Interdisciplinarity (2nd ed.). New York: Oxford University Press.

Hunston, S. (2000). Evaluation and the planes of discourse: Status and value in persuasive texts. In S. Hunston, & G. Thompson (Eds.), Evaluation in text: Authorial stance and the construction of discourse (pp. 177–206). Oxford: Oxford University Press.

Hunston, S. (2011). Corpus approaches to evaluation: Phraseology and evaluative language. London: Routledge.

Klein, J. T. (2015). Interdisciplining Digital Humanities: Boundary Work in an Emerging Field. Michigan: University of Michigan Press.

Klein, J. (2017). A taxonomy of interdisciplinarity. In R. Frodeman, J. Klein, & R. Pacheco (Eds.), The Oxford Handbook of Interdisciplinarity (2nd ed., pp. 21-34). Oxford: Oxford University Press.

Muguiro, N. (2020). Citations in Interdisciplinary Research Articles. Cambridge: Cambridge University Press.

Muguiro, N. (2022). Academic Values in Interdisciplinary Research Articles: A Case Study of Adjectives of Importance. In L. Buckingham, J. Dong, & F. Jiang (Eds.), Interdisciplinary Practices in Academia. Writing, Teaching and Assessment (pp. 51-75). New York: Routledge.

Oakey, D., & Russell, D. (2014). Beyond single domains: Writing boundary crossing. In E. Jakobs, & D. Perrin (Eds.), Handbook of writing and text production (pp. 385-411). Berlin: Mouton de Gruyter.

Spinaci, G., Colavizza, G., & Peroni, S. (2022). A map of Digital Humanities research across bibliographic data sources. Digital Scholarship in the Humanities, 37(4), 1254-1268.

Svensson, P. (2010). The Landscape of Digital Humanities. Digital Humanities Quarterly, 4(1), 1-38.

Thompson, P., & Hunston, S. (2020). Interdisciplinary Research Discourse: Corpus Investigations into Environment Journals. London: Routledge.
Thompson, P., Hunston, S., Murakami, A., & Vajn, D. (2017). Multi-Dimensional Analysis, text constellations, and interdisciplinary discourse. International Journal of Corpus Linguistics, 22(2), 153-186.

Gendered worlds in contemporary Young Adult fiction

Sarah Mukherjee, **Sally Hunt**, **Maria Leedham**

The Open University

Young Adult fiction (YAF), aimed at 11-18 year olds, has a significant role to play in the socialisation of young readers, representing the world and its people, and creating expectations and explanations of their place within it. To date, most YAF research has focused on the themes and styles of individual books, authors or book series with little attention paid to the sex of authors across genres (exceptions are Cermakova and Farová, 2017; Hunt, 2025). In this paper we consider whether female and male authors construct differing worlds with different emphases and preoccupations. The dataset for this study is a newly-compiled, 5-million-word corpus of the 50 most commercially-successful YAF in the UK over a 5-year period (2017-2022, Exploring Discourses of Representation in young Adult fiction [DoRA], Leedham et. al., 2024). The corpus was divided into DoRA_F (34 books) and DoRA_M (16 books) and explored through keyword analysis using AntConc (Anthony, 2024) and subsequent thematic categorisation to highlight the different ways in which authors construct their fictional worlds. Initial findings were investigated in one secondary school through student focus groups and librarian interviews (n=18), providing insightful responses which enriched our subsequent investigations. The study revealed that female and male authors construct starkly distinct worlds with female authors focusing on characters' embodied emotions (e.g. *my smile **fades**, he just **shrugs***), whereas male authors focus on characters' interactions with the natural world (e.g. *Aaron thrashing in the **muck**, he **crawled** through the hole,*). This contrast between a focus on internal and external worlds appears to be a distinctive area of difference between female and male authors of YAF, and cuts across genres and subject matter. We argue that gendered worlds are a fundamental, yet often overlooked, aspect of YAF which have a significant impact in the building of young people's worldviews.

List of references

Anthony, L. (2024). AntConc (Version 4.3.1) [Computer Software]. Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software/AntConc

Cermakova, Anna and Farová, Lenka (2017) His eyes narrowed — her eyes downcast: contrastive corpus-stylistic analysis of female and male writing. Linguistica Pragensia 27(2):7

Hunt, Sally (2025) Linguistic representations of Gender in Children's Literature: Feeling, Speaking, Doing Palgrave Macmillan.

Leedham, M., Hunt S., and Mukherjee, S.J. (2024) Uncovering Discourses of Representation in Young Adult Fiction (DoRA). The Open University. https://wels.open.ac.uk/research/projects/young-adult-fiction-project. Accessed 06 January 2025.

A collocation analysis of student translations in English-Turkish MUST corpus

Meltem Muşlu[1,2]

[1]Gaziantep University; [2]Univerité catholique de Louvain

As Granger and Lefer (2020) argue, the analysis of translations produced by learners can be highly valuable for both pedagogical and research purposes. For instance, learner translations can provide invaluable insights into the understanding of interlanguage development. Collocations have been studied widely in learner corpus research (e.g., Durrant and Schmitt, 2009; Nesselhauf, 2003). However, as pointed out by Vaičenonienė (2023), they have received little attention in corpus-based translation (notable exceptions include Ferraresi and Bernardini, 2022). This study aims to analyze the translations of selected collocations so as to uncover traces of source-language interference or stylistic simplification and the translation decisions in student translations. The data is gathered from Multilingual Student Translation (MUST) corpus (Granger and Lefer, 2020). The MUST English-to-Turkish sub-corpus consists of 365 student translations (88,558 words) produced by undergraduate foreign language learners in Turkiye, with little to no translation training. The source text is a news item about surfing. The data was error-tagged using the Translation-oriented Annotation System (Granger and Lefer, 2021). Six erroneous collocations were selected (corresponding to 2,190 translation solutions) and categorized into the following three translation categories: (1) two collocations that can be translated literally in Turkish, while maintaining the source-text meaning (grab the board, spin the board), (2) two collocations that cannot be translated literally because of the source-text context (takeoff zone, annoyingly infrequent), and (3) two collocations that cannot be translated literally (hit the ocean, package tourists). The results of the study showed that the student translations of collocations display source-language influence and a variety of stylistic choices, sometimes with less successful translation solutions. The results also showed that the most frequent mistakes the students made were related to the verbal constructions and the context. The pedagogical implications of these findings for translator trainers and foreign language educators will be discussed.

Acknowledgement

This research is part of a project that was funded by TUBITAK 2219 International Postdoctoral Research Fellowship Program.

List of references

Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? International Review of Applied Linguistics, 47, pp. 157–177. https://doi.org/10.1515/iral.2009.007

Ferraresi, A. & Bernardini, S (2022). Comparing collocations in translated and learner language. International Journal of Learner Corpus Research 9:1 (2023), pp. 125–153. https://doi.org/10.1075/ijlcr.22012.fer

Granger, S. & Lefer, M.A. (2021). Translation-oriented Annotation System manual (Version 2.0). CECL Papers 3. Louvain-la-Neuve: Centre for English Corpus Linguistics/Université catholique de Louvain. https://cdn.uclouvain.be/groups/cms-editors-cecl/cecl-papers/TAS-2.0_annotation_manual_2021-10-26.pdf

Granger, S., & Lefer, M.A. (2020). The Multilingual student translation corpus: a resource for translation teaching and research. Language Resources and Evaluation, 54, 1183-1199. https://doi.org/10.1007/s10579-020-09485-6

Nesselhauf, N. (2003). The Use of Collocations by Advanced Learners of English and Some Implications for Teaching. Applied Linguistics, 24(2), pp. 223–242. https://doi.org/10.1093/applin/24.2.223

Vaičenonienė, J., (2023). Besimokančiųjų vertimo studentų tekstynas MUST-LT: kolokacijų vertimo atvejo analizė (Student Translation Corpus MUST-LT: a case analysis of collocation translation). Studies about Languages / Kalbų studijos, 42, 105-118.

"It's not unusual": Double negation in academic lectures

Hilary Nesi[1], Benet Vincent[1], Betül Bal Gezegin[2]

[1]Coventry University; [2]Ondokuz Mayıs University

While negated propositions require more complex cognitive processing than affirmative propositions, double negation(DN; exemplified by it's not unusual) is even more difficult to make sense of in English, especially as it can be used in the service of a variety of different stance positions. Thissophistication can contribute to misunderstandings by speakers of other languages, especially if DN is used differently in their first languages. However, the use of DN in university lectures seems not to have been researched. Ourstudy investigates the use of DN in academic lectures in English, comparing its occurrence in two contrasting contexts, institutions in English-speaking countries, and EMI (English as a Medium of Instruction) institutions. The first context is represented by the British Academic Spoken English (BASE)corpus, the Michigan Corpus of Academic Spoken English(MICASE) and a corpus of open courseware lectures from the US (OCC). The EMI lectures we consult come from EmiBO(Johnson & Picciuolo 2022) from the University of Bologna, acorpus of lectures delivered by Turkish lecturers in Turkey(TEMI), and CoFEL, a corpus of French Engineering Lectures (Picavet 2024). We used CQL queries on Sketch Engine to retrieve examples of DNs and checked these to identify relevant hits. This showed that DN is not infrequently found in corpora of lectures from Anglophone countries, but appearsvanishingly rare in the EMI context. Participants in EMI lectures may therefore lack exposure to DNs and their range of expression. Following further qualitative analysis we have created a framework of DN functions in academic discourse. Our talk will introduce the framework and discuss examples showing how lecturers employ DN in various ways. Our study can help raise educators' awareness of this rather neglected linguistic feature, and lead to a deeper understanding of variation in academic discourse across cultural and instructional settings.

List of references

Johnson, J.H., & Picciuolo, M. (2022). THE EMIBO CORPUS A resource for investigating lecture discourse across disciplines and lecture modes in an EMI context. Lingue e Linguaggi, 53, 253-272.

Picavet, F. (2024). Summarizing and storytelling in English-medium engineering education [Doctoral thesis]. Université Grenoble Alpes (UGA).

Lexical and stylistic characteristics of helpful and unhelpful book reviews

Anastasia Novoselova

University of Wolverhampton

Online product reviews have become an important consideration in people's purchase decisions. The aim of the study is to obtain insights into the relationship between discursive content of reviews and consumer perception of their helpfulness.

We constructed a large corpus of online book reviews that have been published on the Amazon website, 1996 - 2023. Sorting the initial collection of reviews by their helpfulness scores, we created a 5-million-word subcorpus from reviews with the greatest number of helpful votes and a 5-million-word subcorpus from reviews with the smallest number of helpful votes. Thus, the former subcorpus contains 19k reviews about 12k books, written by 16k unique reviewers, and the latter subcorpus contains 131k reviews about 89k books, written by 71k unique authors.

Using corpus-assisted discourse analysis methods (Baker, 2020; Partington et al., 2013), we identified stylistic and lexical means associated with either type of reviews. First, consistent with a number of previous studies on the topic (Mudambi and Shuff, 2010; Pollach, 2006; Wu et al., 2011), we find that the mean length of helpful reviews (262.2) is greater than the mean of unhelpful ones (38.1) to a statistically significant degree (according to an independent samples t-test); the type-token ratio of helpful reviews is similarly significantly greater than that of unhelpful reviews (1.88% vs 1.65%). Secondly, we identify lexical distinctions of the two subcorpora. For example, helpful book reviews tend to be characterised by a greater number of self-mentions of a reviewer; they more often include background information about the reviewer; they explicitly state the purpose of the review and justify the star-rating; helpful reviews tend to address the reader, and mention likely preferences, habits and identities of those readers who are likely to enjoy or dislike the book (Chik and Taboada, 2020; Skalicky, 2013; Virtanen, 2017; Zou and Hyland, 2022).

List of references

Baker, P. (2020) 'Corpus-assisted discourse analysis' in C. Hart (ed.) Researching Discourse: a student guide. London: Routledge, pp.124-142.

Chik, S. and Taboada, M. (2020) Generic structure and rhetorical relations of online book reviews in English, Japanese and Chinese. Contrastive Pragmatics, 1, pp.143-179.

Mudambi, S. and Shuff, D. (2010) What makes a helpful online review? A study of customer reviews on Amazon.com. MIS Quarterly, 34 (1), pp.185-200.

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

Pollach, I. (2006). Electronic word of mouth: a genre analysis of product reviews on consumer opinion web sites. System Sciences, 2006. HICSS'06. Proceedings of the 39th Hawaii International Conference on System Sciences. Volume 3.

Skalicky, S. (2013) Was this analysis helpful? A genre analysis of the Amazon. com discourse community and its "most helpful" product reviews. Discourse, Context & Media 2(2), pp. 84–93.

Virtanen, T. (2017) Adaptability in online consumer reviews: exploring genre dynamics and interactional choices. Journal of Pragmatics, 116, pp.77-90.

Wu, Ph., Heijden, H. and Korfiatis, N. (2011) The influence of negativity and review quality on the helpfulness of online reviews. Proceedings of the 32nd International Conference on Information Systems. Shanghai.

Zou, H. and Hyland, K. (2022). How the medium shapes the message: stance in two forms of book reviews. Journal of Pragmatics, 193, 269-280.

A comparative analysis of verb usage in the National Corpus of Irish and the corpus of spoken Irish

Mícheál J. Ó Meachair, Gearóid Ó Cleircín, Kevin Scannell, Brian Ó Raghallaigh, Úna Bhreathnach

Dublin City University

In this paper we draw on the newly available tools of the National Corpus of Irish (Corpas Náisiúnta na Gaeilge, CNG) and the Corpus of Spoken Irish (Corpas na Gaeilge Labhartha, CGL) at corpas.ie to analyse verb use in written and spoken contexts. CNG contains 101 million words, including a wide variety of text types and genres, and is balanced in a way that is representative of the Irish-language for the years 2000 to 2024. CGL contains 9 million words, and is a specialized corpus that includes transcriptions from the radio, television, podcasts and YouTube. CGL also includes subtitles and scripts for television, scripts for plays or pantomimes, song lyrics, prayers, and similar data that were written to be read or said aloud.

Our analysis begins with an overview of all conjugated verb forms in CNG and CGL. These conjugated verb forms include: the past and past continuous, the present and present continuous, and the future tense. We go on to explore the distribution of conjugated verb forms across genres in each corpus. We conclude this section of our analysis with explanations of why, where possible, certain verbs are used more frequently than others in a given context. These results are presented in a manner that aims to support second-language acquisition (SLA) research in Ireland. Finally, this work serves as a technical and methodological test of the newly created corpus analysis tools on corpas.ie.

List of references

(forthcoming - we have read quite broadly for this paper, and we are refining our references as we write up the journal article)

A corpus-informed investigation of Irish dialect in Rudyard Kipling's "Three Musketeers" stories

David Oakey[1], **Jenny Amos**[2]

[1]University of Liverpool; [2]University of Suffolk

This paper presents a corpus-informed dialectal examination of Rudyard Kipling's series of short stories featuring three soldiers stationed in India during the British Raj, known as the "Three Musketeers". Each speaker's contribution is written with orthographic alterations indicating differing dialects - Ortheris, a Cockney; Learoyd, a Yorkshireman; Mulvaney, an Irishman. In the 140 years since these stories were published, first in India, then, internationally, Kipling's use of dialect writing has been both praised and vilified by literary critics. An anonymous reviewer in 1899 found that "Mr. Kipling has a genius for reproducing quaint and characteristic Hibernicisms" (p42), and Raine (1992, p12) argued that "dialect is Kipling's greatest contribution to modern literature … and he is the most accomplished practitioner since Burns." Others disagree. Buchanan (1899/1971), for example, referred to Kipling's Irish character as "speaking a dialect which would cause amazement in the Emerald Isle" (p242), and Dudley Edwards thought little of "Kipling's laboured attempts in prose to reproduce Irish dialect" (Dudley Edwards 1988, p3).

To our knowledge, however, there has been no systematic attempt to describe the Irish dialect found in Kipling's *Three Musketeers* stories which would help to support or reject the opinions of these many critics. This paper, therefore, uses a corpus of the stories, and AntConc (Anthony 2024) and Sketchengine (Kilgariff et al 2014) software to investigate Kipling's dialect writing, focusing on the direct speech of Mulvaney. We first discuss the well-known techniques and difficulties in expressing any dialect in writing, particularly in the absence of a prescribed standard for Irish (Connell 2014 p39), before reviewing notable phonological features of Irish English (Kallen 1994; Hickey 2004), and how these are represented across the stories in the corpus. Features include NEAR to SQUARE (e.g. 'quare' for *queer*), DRESS to KIT (e.g. 'whin' for *when*), and, orthographic, 'th' to 'd' (e.g. 'wid' for *with*).

List of references

•Anonymous. (1899/1971). 'Review of Soldiers Three', The Spectator, Vol. LXII, pp. 403-4, 23 March 1889, reprinted in the Kipling Journal, Vol. VII, No. 52, pp. 27—30, December 1939. In R. Lancelyn Green (Ed.), (1971) Rudyard Kipling: The Critical Heritage (pp. 41-43). London: Routledge.
•Anthony, L. (2024). AntConc 4.3.1. https://www.laurenceanthony.net/software/antconc/.
•Buchanan, R. (1899/1971). 'The Voice of the Hooligan', Contemporary Review, Vol. LXXVI pp. 774-89. In R. Lancelyn Green (Ed.), Rudyard Kipling: The Critical Heritage (pp. 233-249). London: Routledge.
•Connell, M. A. (2014). It's in the details the divil is: Corpus linguistics and Irish English literary dialect. Unpublished PhD Thesis, Department of English, National University of Ireland, Galway.
•Dudley Edwards, O. (1988). Kipling and the Irish. London Review of Books, 10(3), 1-10. Retrieved 16th August 2024 from https://www.lrb.co.uk/the-paper/v10/n03/owen-dudley-edwards/kipling-and-the-irish.
•Kallen, J. L. (1994). English in Ireland. In R. Burchfield (Ed.), The Cambridge History of the English Language: Volume 5: English in Britain and Overseas: Origins and Development (Vol. 5, pp. 148-196). Cambridge: Cambridge University Press.
•Hickey, R. (2004). The phonology of Irish English. In B. Kortmann & E. W. Schneider (Eds.), Handbook of varieties of English. Volume 1: Phonology (pp. 68-97). Berlin: Mouton de Gruyter.
•Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., . . . Suchomel, V. (2014). The Sketch Engine: ten years on. Lexicography, 1(1), 7-36.
•Raine, C. (1992). Kipling and Modernism. London Review of Books, 14(15), 1-12. Retrieved 9th January 2025 from https://www.lrb.co.uk/the-paper/v14/n15/craig-raine/kipling-and-modernism.

Effects of evidence-based tagset development on POS tagging accuracy

Vlasta Ohlídalová[1,2]

[1]Masaryk University; [2]Lexical Computing

After years of almost stagnation, the Part-of-Speech tagging for morphologically rich languages has seen a significant improvement with the rise of neural networks. Despite the relatively high accuracy the state-of-the-art models achieve (currently above 97 % for Czech language), it still means that on average, there is a mistake in almost every second sentence and even the best models still frequently fail us on the phenomena that we are most interested in (e.g. the difference between nominative and accusative case that often denotes the distinction between object and subject).

While creating new algorithms is the commonly used method of increasing the accuracy, this paper described an alternative strategy - that is to investigate how the tagging accuracy is influenced by modifications in the dataset. As the borders between some POS categories are blurry and agreement on various language phenomena is rather low, a large scope opens up for possible modifications.

In the paper, we describe numerous experiments we performed using linguistically motivated modifications with the goal of achieving better accuracy while keeping the same tool.

The experiments were conducted for the Czech language, utilizing the RFTagger for evaluation purposes. Due to the tagger's efficiency, we were able to conduct as many experiments as needed and evaluate results on manually annotated corpus DESAM using the 10-fold cross-validation.

Ultimately, five of the experiments resulted in a significant increase in accuracy. When these results were combined, the final dataset achieved a reduction in the error rate of approximately 12% (from 7.1% to 6.3%).

While the final accuracy does not reach state-of-the-art results, it demonstrates that this approach is a viable alternative for enhancing the accuracy of Part-of-Speech tagging.

List of references

OHLÍDALOVÁ, Vlasta. Improvements of the tagset used for automatic morphological analysis of Czech. Online. Diplomová práce. Brno: Masarykova univerzita, Filozofická fakulta. 2023.

Karel Pala, P. Rychlý, and Pavel Smrz. DESAM - Annotated Corpus for Czech. In: Conference on Current Trends in Theory and Practice of Informatics. 1997

Helmut Schmid and Florian Laws: Estimation of Conditional Probabilities with Decision Trees and an Application to Fine-Grained POS Tagging, COLING 2008, Manchester, Great Britain.

A haven of hope: Positive representation of Ukrainian refugees in the UK and Spanish press

Aroa Orrequia-Barea

Universidad de Cádiz

This research is part of a larger project that examines media representation of Ukrainian refugees in the UK and Spanish press, comparing it with the discursive construction of other migrant groups in the press (Baker & McEnery 2005, Gabrielatos & Baker 2008, Baker et al. 2008, KhosraviNik 2009, 2010a, 2010b). In this context, two research questions are posed: (1) How are Ukrainian refugees discursively constructed in the media?; and (2) is there any difference in the representation of other migrant groups?

The corpora consist of the news items published by selected newspapers in the UK and Spain, compiled during the first month of the war (24th February to 24th March, 2022). The software *Sketch Engine* was used to analyse the texts.

According to previous research, there is a tendency for migrants to receive negative-other representation in the press (Baker, Gabrielatos & McEnery, 2013). However, the results of this study indicate that Ukrainians have been positively represented in both countries. While there is a preference for the term “refugee”/"refugiado" in the case of the Ukrainians, other migrants are often referred to as “asylum seekers” and/or “immigrants”/"immigrantes", which carry negative connotations. In fact, “asylum”/"asilo" is not even among the most frequent words in these corpora despite being a part of the process (Gabrielatos & Baker, 2008, p. 17). Additionally, water metaphors and large-quantity words, which have traditionally been used to de-humanise migrants (Taylor, 2021), are employed positively to emphasise the urgency for the humanitarian crisis.

Similarly, Ukrainians are individualised through the use of their proper names, ages, character building and direct quotes; whereas other migrants are collectivised. These discursive strategies contribute to a sympathetic representation of Ukrainian refugees, in contrast to the negative other-presentation of other migrants. Treating people as human beings should be the general norm, not the exception.

List of references

Baker, P., & McEnery, T. (2005). A corpus-based approach to discourses of refugees and asylum seekers in UN and newspaper texts. Journal of language and politics, 4(2), 197-226.

Baker, P., Gabrielatos, C. & McEnery, T. (2013). Analysis and Media Attitudes: The Representation of Islam in the British Press. Cambridge.

Gabrielatos, C., & Baker, P. (2008). Fleeing, sneaking, flooding: A corpus analysis of discursive constructions of refugees and asylum seekers in the UK press, 1996-2005. Journal of English linguistics, 36(1), 5-38. https://doi.org/10.1177/0075424207311247

KhosraviNik, M. (2009). The representation of refugees, asylum seekers and immigrants in British newspapers during the Balkan conflict (1999) and the British general election (2005). Discourse & Society, 20(4), 477-498. https://doi.org/10.1177/0957926509104024

KhosraviNik, M. (2010a). The representation of refugees, asylum seekers and immigrants in British newspapers: A critical discourse analysis. Journal of language and Politics, 9(1), 1-28. https://doi.org/10.1075/jlp.9.1.01kho

KhosraviNik, M. (2010b). Actor descriptions, action attributions, and argumentation: Towards a systematization of CDA analytical categories in the representation of social groups. Critical discourse studies, 7(1), 55-72. https://doi.org/10.1080/17405900903453948

Taylor, C. (2021). Metaphors of migration over time. Discourse & Society, 32(4): 463-481. https://doi.org/10.1177/0957926521992156

Exploring p-frames in university students' talk: Variation, predictability, and discourse functions

Yusuf Ozturk[1], Federica Barbieri[2]

[1]Mus Alparslan University; [2]Swansea University

Most research on formulaic language to date has focused on continuous sequences of three or more words (e.g., *I don't know what*), typically referred to as lexical bundles or n-grams. However, there is also a long-standing interest in discontinuous sequences of words 'in which words collocate at a distance or in varying sequences' (Gray & Biber, 2013, p. 110). P-frames are discontinuous patterns containing one or more variable slots in a sequence of words (e.g. *it is * to*, with variants *interesting, useful, nice*) (Gray & Biber, 2013). Because they carry information about pattern variability (Biber, 2009; Römer, 2010), p-frames can provide valuable information on the phraseological competence of speakers. However, to date p-frames have been investigated mostly in writing, and the limited research on p-frames in speech (e.g., Nekrasova-Beker, 2021) relied on elicited data from language tests.

The present study examines p-frames in university students' talk in English. Specifically, it focuses on three key properties of p-frames: variability (i.e. variant/p-frame ratio), predictability (i.e. the extent to which variants are evenly/unevenly distributed, shown by entropy values), and discourse function (Biber et al., 2004). The study is based on three corpora of informal student talk, representing three types of university students: British and international students at a UK university, and Turkish students at a Turkish university. The corpora comprise transcribed semi-structured interviews with students (about 50 per group) in these contexts, and comprise ca. 140,000, 87,000, and 83,000 words respectively. AntConc (Anthony, 2024) was used to extract the 100 most frequent 4-word p-frames (with at least two medial variants). Preliminary findings suggest that British and international students use more unpredictable p-frames, while Turkish students' p-frames are more fixed. The p-frames retrieved will also be compared to those identified in the conversation register by Gray and Biber (2013).

List of references

Anthony, L. (2024). AntConc (Version 4.3.1) [Computer Software]. Waseda University. https://www.laurenceanthony.net/software/AntConc

Biber, D. (2009). A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. International Journal of Corpus Linguistics, 14(3). https://doi.org/10.1075/ijcl.14.3.08bib

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at …: Lexical Bundles in University Teaching and Textbooks. Applied Linguistics, 25(3). https://doi.org/10.1093/applin/25.3.371

Gray, B., & Biber, D. (2013). Lexical frames in academic prose and conversation. International Journal of Corpus Linguistics, 18(1), 109-136.

Nekrasova-Beker, T. (2021). Use of phrase-frames in L2 students' Oral production Across proficiency sub-Levels. In C. William (Ed.), Multiple Perspectives on Learner Interaction (pp. 41-68). De Gruyter Mouton. https://doi.org/doi:10.1515/9781501511370-004

Römer, U. (2010). Establishing the phraseological profile of a text type: The construction of meaning in academic book reviews. English text construction, 3(1), 95-119.

A multilingual exploration of environmental stakeholders' discourse: Insights from an NGO's report

Adriana Silvina Pagano[1], Cristina Bosco[2], Elisa Chierchiello[2], Irena Rašin[3], Olena Kozan[4], Stella Markantonatou[5]

[1]Universidade Federal de Minas Gerais; [2]Università di Torino; [3]Sveučilište u Zagrebu; [4]ANKARA HACI BAYRAM VELİ ÜNİVERSİTESİ; [5]Athena Research Center

Within a broader project aimed at exploring how stakeholders communicate ecological vulnerability and following the current trend of compiling corpora to be used in NLP for resourcing actions promoting biodiversity and sustainability (see e.g. Grasso et al., 2024), this paper reports results from a study of a multilingual corpus compiling World Wide Fund for Nature (WWF)'s 2024 report in English and its human translations into Italian, Spanish, Brazilian Portuguese, Turkish, Croatian and Greek. Drawing on systemic-functional theory (Halliday & Matthiessen, 2014), our analysis aimed to explore how meanings were construed in each language considering the meta-context of translation and the different environments at which translation equivalence was established (Matthiessen, 2001). We retrieved source and target texts from WWF's websites and first compared them to assess the extent of translated content and images in the target versions. Subsequently, we converted and manually inspected the pdf files into text and tokenised them into sentences to semi-automatically align source and target equivalents. We obtained raw word frequency, type/token ratio, clusters and collocations using corpus tools and performed sentiment and topic analysis using Natural Language Processing (NLP) techniques. We leveraged corpus data to explore systemic-functional categories. Hence, we examined the top ten ranking nouns in terms of denotation and connotation, agency (human vs. non-human) and level of abstraction (abstract vs. concrete). We explored verbs in terms of types of experiential meaning construed. Our results point to the impact of translation on distinct perceptions of key themes in environmental discourse by different language communities. Our corpus will be made publicly available and is expected to contribute to building NLP resources to explore the language of environmental communication.

List of references

Grasso, Francesca, et al. "EcoVerse: An Annotated Twitter Dataset for Eco-Relevance Classification, Environmental Impact Analysis, and Stance Detection." arXiv preprint arXiv:2404.05133 (2024).

Halliday, Michael Alexander Kirkwood, and Christian MIM Matthiessen. Halliday's introduction to functional grammar. Routledge, 2013.

Matthiessen, C. M. I. M. "The environments of translation." Exploring translation and multilingual text production: Beyond content (2001): 41-124.

Stede, Manfred, and Ronny Patz. "The climate change debate and natural language processing." Proceedings of the 1st Workshop on NLP for Positive Impact. 2021.

Decoding deception: A corpus-assisted discourse analysis of COVID-19 phishing strategies

Nicola Pelizzari

University of Brescia

Phishing, a form of cybercrime that exploits trust to acquire sensitive information fraudulently (Jakobsson & Myers, 2007), intensified during the COVID-19 pandemic, with complaints related to phishing surging by 69% as COVID-themed scams exploited societal fears and uncertainties (Europol, 2021). To examine how these scams operated linguistically, this study employs corpus-assisted discourse analysis (CADS) to investigate 1,609 phishing emails (465,446 words) collected in the UK between 2020 and 2023. The corpus, filtered for pandemic-related content and organised diachronically into yearly subcorpora, enables an exploration of how phishing language reflected or responded to socio-political shifts such as lockdowns, vaccine rollouts, and economic aid initiatives. By integrating quantitative corpus techniques with qualitative discourse analysis (Baker et al., 2008; McEnery & Hardie, 2012), this study traces how linguistic strategies evolved in response to shifting socio-political contexts. Three dominant rhetorical tactics are identified: (1) urgency, marked by temporal markers and appeals to scarcity; (2) impersonation of authoritative entities through institutional mimicry; and (3) exploitation of fear and hope, leveraging pandemic-specific keywords and emotional triggers. Diachronic analysis reveals significant linguistic shifts: health-related themes peaked during lockdowns and vaccine campaigns (2020–2021), while financial narratives resurged post-2022 as pandemic concerns declined. Statistical validation ($\chi^2 = 1604.83$, $p < .001$) confirms these trends, illustrating phishing's adaptability to global crises (Canfield et al., 2016; Ferreira & Teles, 2019). These findings highlight the value of CADS in cybersecurity research, not only for revealing how phishing strategies adapt to global events but also for providing actionable insights for enhancing email filtering systems and user education programs. They underscores the need for dynamic, context-aware preventative measures (Hutchings & Clayton, 2016).

List of references

Baker, P., Gabrielatos, C., KhosraviNik, M., Krzyzanowski, M., McEnery, T., & 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 & Society, 19(3), 273–306.

Canfield, C. I., Fischhoff, B., & Davis, A. (2016). Quantifying phishing susceptibility for detection and behavior decisions. Human Factors, 58(8), 1158–1172.

Europol. (2021). Internet organised crime threat assessment (IOCTA) 2021. Retrieved from https://www.europol.europa.eu/publications-events/main-reports/internet-organised-crime-threat-assessment-iocta-2021 (Last accessed on January 2, 2025).

Ferreira, A., & Teles, S. (2019). Persuasion: How phishing emails can influence users and bypass security measures. International Journal of Human-Computer Studies, 125, 19–31.

Hutchings, A., & Clayton, R. (2016). Exploring the provision of online booter services. Deviant Behavior, 37(10), 1163–1178.

Jakobsson, M., & Myers, S. (2007). Phishing and countermeasures: Understanding the increasing problem of electronic identity theft. John Wiley & Sons.

McEnery, T., & Hardie, A. (2012). Corpus linguistics: Method, theory and practice. Cambridge University Press.

Verbs in context: Using BERT to study grammar patterns

Florent Perek

University of Birmingham

Distributional semantics is increasingly used as a research tool in many disciplines of linguistics, notably to study lexico-grammatical associations (e.g. Perek 2018). However, current approaches tend to draw on type-based distributional semantic models (e.g. word2vec), in which a single representation for each word is derived from all its instances in the training corpus, thus ignoring homonymy and polysemy. This presents a limitation for grammar studies, as grammatical constructions are typically sensitive to the meaning of lexical items or modulate them in various ways. It has been suggested that token-based models such as BERT (Devlin et al. 2019), which provide semantic representations for each instance of a word depending on the context, can offer a solution to the limitations of type-based models.

This paper explores the suitability of BERT to study the meaning of verbs in context, using data from the COBUILD Grammar Patterns (Francis et al. 1996). Lists of verbs attested in three patterns ("V n for n", "V n with n", and "V for n") were taken from COBUILD, and examples of these verbs in the corresponding patterns were extracted from the BNC using a dependency parser. These examples were then submitted to BERT, creating token vectors, i.e. semantic representations of every verb token in context. Four main findings emerge from the analysis of this data:

1. The clusters obtained by grouping token vectors are very similar to the meaning groups from COBUILD, which are precisely based on the meaning of verbs in the pattern.

2. BERT also uncovers meaningful groupings that are not mentioned in COBUILD.

3. BERT can detect different senses of the same verb in the pattern, and sort them with similar meanings.

4. In many cases, BERT can semantically distinguish instances of a verb in a pattern from instances of the same verb in other constructions.

List of references

Devlin, J., Chang, M. W., Lee, K. and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 4171–4186.

Francis, G., Hunston, S. and Manning, E. (1996). Collins COBUILD Grammar Patterns 1: Verbs. London: HarperCollins.

Perek, F. (2018). Recent change in the productivity and schematicity of the way-construction: A distributional semantic analysis. Corpus Linguistics and Linguistic Theory 14(1): 65–97.

The role of academic vocabulary in shaping Sustainable Development Goal (SDG) research papers: A lexical multidimensional perspective

Paula Pinto[1], Tony Berber Sardinha[2], Denis Owa[3], Maria Claudia Delfino[4], Simone Resende[5]

[1]Sao Paulo State University (UNESP); [2]Pontifícia Universidade Católica de São Paulo (PUC/SP); [3]Pontifícia Universidade Católica de São Paulo (PUC/SP); [4]Faculdade de Tecnologia de São Paulo (FATEC); [5]Faculdade Cultura Inglesa

Since the establishment of the United Nations Agenda 2030, researchers across the globe have increasingly addressed complex challenges - such as hunger, poverty, education, and sustainability - within the framework of the Sustainable Development Goals (SDGs). During the COVID-19 pandemic, international collaboration intensified, prompting a surge of publications that framed research through the lens of the SDGs. These publications, often authored in English and disseminated in high-impact journals, present significant linguistic and stylistic challenges for non-English-speaking scholars aiming to contribute to global academic discourse. Familiarity with the vocabulary used in such contexts is therefore essential - not only to understand prevailing research trends but also to produce work that aligns with the expectations of the international academic community. This study investigates the lexico-grammatical patterns in English-language SDG-related research articles by conducting a Lexical Multidimensional Analysis of a 2-million-word corpus compiled from the PLOS platform using the term “Sustainable Development Goal” (Pinto, 2021; 2023). The corpus was part-of-speech tagged with TreeTagger in Sketch Engine, and six dimensions of lexical variation were identified: Government Actions, Presenting Results, Data Interpretation, Data Presentation, Data Quality, and Research Procedure. While the first dimension reflects content specifically related to the SDGs, the remaining five capture the recurrent discourse strategies researchers use to describe methodologies and findings. Building on these insights, we developed academic writing workshops for Brazilian researchers. These workshops use data-driven activities to support the production of research articles and foster critical reflection on the dominant lexis and structures present in SDG-focused publications. By connecting corpus-driven findings with pedagogical practice, this initiative contributes to the internationalization of academic writing practices among Brazilian scholars.

Acknowledgement
The authors acknowledge the financial support of the following organizations: National Council for Scientific and Technological Development (CNPq)Grant #307287/2021-1/São Paulo Research Foundation (FAPESP), Grant #2022/05848-7.

List of references

Berber Sardinha, T. (2022). A text typology of social media. Register Studies, 4(2), 138-170.

Delfino, M. C., Araújo, R. F., & Berber Sardinha, T. (2018). Revista brasileira de Linguística Aplicada: multidimensões temáticas. In M. J. B. Finatto, R. R. Rebechi, S. Sarmento, & A. E. P. Bocorny (Eds.), Linguística de Corpus: Perspectivas (pp. 93-124). Porto Alegre: Instituto de Letras da UFRGS.

Pinto, P. T. (2021). The Sustainable Development Goals (SDGs) words ‘poverty’ and ‘sustainability’ in Brazilian research: a preliminary thematic corpus-based analysis. Cadernos De Linguística, 2(4), e440. https://doi.org/10.25189/2675-4916.2021.v2.n4.id440

Pinto, P. T., Serpa, T., Silva, E. B. da, & Camargo, D. C. de. (2023). Quando o léxico geral se torna terminologia no contexto social: Um estudo sobre os termos relacionados aos Objetivos de Desenvolvimento Sustentável (ODS) das Nações Unidas. Letras & Letras, 39, e3905 | p. 1–31. https://doi.org/10.14393/LL63-v39-2023-05

Sardinha, T. B. (2020). A historical characterisation of American and Brazilian cultures based on lexical representations. Corpora, 15(2), 183-212.

Pushing the boundaries: Creating a Danish semantic tagger for metaphor analysis of cancer narratives

Sander Puts[1], Daisy Lal[2], Mette Byg Josefsen[3], Bettina Mølri Knudsen[3], Kasper Frank[3], Paul Rayson[2], Leonard Wee[1]

[1]Maastro, Netherlands; [2]UCREL, Lancaster University; [3]Center for Shared Decision Making, Lillebaelt Hospital

In health communication in general, and when people talk about cancer or end of life care in particular, metaphor is an important phenomenon to consider. In previous work, the USAS semantic analysis tool (Rayson et al., 2004) was applied as part of English metaphor analysis, where studies showed, for example, that Violence metaphors are not by default negative and Journey metaphors are not by default a positive means of conceptualising cancer (Semino et al, 2017). Now, in the 4D Picture project (https://4dpicture.eu/), a large European multi-disciplinary consortium of researchers is aiming to improve the cancer patient experience and ensure that personal preferences are respected. One part of the project involves the semi automatic analysis of online forums, interview transcripts and patient questionnaire free text answers on a large scale and in four languages (Danish, Dutch, English and Spanish) to investigate the experiences of cancer patients and their significant others via an analysis of themes, values, preferences and concerns and how they relate to treatments and cancer stages. We describe the research undertaken to create a new semantic analysis system for Danish to be incorporated into the open source Python version of the semantic tagger (PyMUSAS). Specific methodological challenges that we have addressed include the creation of high quality linguistic resources in the lexicons where we have used a combination of online sources for multiword expression (MWE) definitions and open-source large language models (LLMs) to perform internet searches to retrieve definitions of MWEs, facilitating accurate translation from English lexicons to Danish. Our evaluations include: a) coverage and accuracy of the lexicons particularly for health domain specific terminology, and b) manual annotation of Danish narratives for metaphor. All resources and data will be released publicly for replication purposes where possible with Creative Commons licences, and PyMUSAS is available with an Apache licence.

List of references

Rayson, P., Archer, D., Piao, S. L., McEnery, T. (2004). The UCREL semantic analysis system. In proceedings of the workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks in association with 4th International Conference on Language Resources and Evaluation (LREC 2004), 25th May 2004, Lisbon, Portugal, pp. 7-12.

Semino E, Demjén Z, Demmen J, et al (2017) The online use of Violence and Journey metaphors by patients with cancer, as compared with health professionals: a mixed methods study. BMJ Supportive & Palliative Care 2017;7:60-66.

Creating a corpus of hard-to access school texts: Ethical and methodological insights from the BAWESS corpus design and compilation

Reka R Jablonkai, Gail Forey

University of Bath

Corpus linguistics has played a significant role in enhancing the teaching and learning of discipline-specific varieties of English. While much of the existing research has focused on publicly available texts and higher education student outputs, there is a clear lack of corpora capturing the disciplinary writings of young learners in secondary school contexts (Durrant, 2023; Deignan et al., 2023). To address this gap, the British Academic Written English Secondary School Corpus (BAWESS) is being developed as part of the ESRC-funded Disciplinary literacy and corpus-based pedagogy project (2025-2028).

This paper examines the challenges and innovations involved in designing and compiling the BAWESS corpus, highlighting both methodological and ethical considerations. The corpus building process required the research team to address nuanced ethical challenges to prioritise transparency and respect for the agency of young participants while safeguarding their privacy. The genres and genre families (Gardner & Nesi, 2013) constituting the sampling frame for corpus building were identified based on the analysis of relevant policy documents (e.g. English National Curriculum), exam papers and interviews with secondary school subject teachers. Metadata collection was carefully designed to anonymise sensitive information while preserving key contextual details, such as demographic variables, exam details, and institutional background, to support meaningful corpus analysis.

The BAWESS corpus aims to fill a critical gap in the field of corpus-based educational research by providing insights into the linguistic characteristics of young learners' disciplinary writing and its development in secondary school settings. A framework of guiding principles for addressing challenges in educational corpus development will be proposed to guide researchers working towards developing corpora of potentially sensitive, hard-to-access texts.

This paper contributes to advancing best practices in corpus building and highlights the potential for corpora of young learners' writing to inform evidence-based approaches to disciplinary literacy, curriculum design, and assessment in secondary education.

List of references

Deignan, A., Candarli, D., & Oxley, F. (2023). The linguistic challenge of the transition to secondary school. Routledge.

Durrant, P. (2023). Corpus Linguistics for writing development. Routledge.

Gardner, S. & Nesi, H. (2013). A classification of genre families in university student writing. Applied Linguistics, 34(1), 25-52.

Corpus-based pedagogy to support the development of disciplinary literacy in secondary school subjects

Reka R Jablonkai, Gail Forey, Natalie Cheers

University of Bath

Corpus-based methods have a rich and successful history in exploring the diverse ways language is employed within different disciplines. Much of this investigation has centered around analyzing disciplinary discourses in research papers and university student writing. However, there is a notable scarcity of research examining language use and disciplinary variation in registers at the secondary school level (Deignan et al., 2023; Durrant & Brenchley, 2023). This paper presents findings from a pioneering study that examined small corpora compiled from high-stakes school examination papers in order to develop and design teaching materials to support the exam preparation in secondary school subjects.

High-stakes examinations at the secondary school level in the UK context (e.g. GCSE and A Level), typically rely on written language. Prior investigations into disciplinary literacy skills have urged a more thorough investigation of these exams (Ricketts et al., 2014) and advocated for a more specialized and language-aware approach to enhance disciplinary literacy (Quigley & Coleman, 2021), that is, the teaching and "learning how to read, think about, write, communicate, and use information like each discipline's experts" (Zygouris-Coe, 2012, p. 36). This study focused on examination papers from three school subjects: Geography, Biology, and History. Three corpora were compiled, comprising past exam papers and model answers in these subjects. To pinpoint relevant linguistic features for developing disciplinary literacy, interviews were conducted with subject teachers. Building upon the interview findings, the corpus analysis focused on single words, lexical bundles, discipline-specific prefixes and suffixes, and keywords. The outcomes of the corpus analysis form the basis of collaboratively designing subject-specific materials for exam preparation with subject teachers. Various corpus-based activities will be proposed along with guidelines for crafting corpus-based disciplinary literacy teaching materials. The study will conclude by discussing the applicability of a corpus-based pedagogy for fostering disciplinary literacy in secondary school settings.

List of references

Durrant, P., & Brenchley, M. (2023). Development of noun phrase complexity across genres in children's writing. Applied Linguistics, 44(2), 239–264.

Deignan, A., Candarli, D., & Oxley, F. (2023). The linguistic challenge of the transition to secondary school. Routledge.

Ricketts, J., Sperring, R. & Nation, K. (2014). Educational attainment in poor comprehenders. Frontiers in Psychology, 5, 445-456. DOI: 10.3389/fpsyg.2014.00445

Quigley, A. & Coleman, R. (2021). Improving literacy in secondary schools Guidance Report. Education Endowment Foundation. Available online: https://educationendowmentfoundation.org.uk/education-evidence/guidance-reports/literacy-ks3-ks4

Zygouris-Coe, V. (2012). Disciplinary literacy and the common core state standards. Topics in Language Disorders, 32(1), 35-50.

Comparative corpus-assisted discourse analysis of the role of experts in news articles about the 1918 influenza and COVID-19 pandemics

Jenni Maria Räikkönen

University of Helsinki

Experts play a central role during health crises, advising both policymakers and the public. However, their visibility and the framing of their statements depend heavily on editorial choices. This study explores how expert voices are represented in pandemic-related news through a comparative, corpus-assisted discourse analysis (CADS).

Two comparisons are made: (1) a diachronic analysis of New York Times (NYT) coverage of the 1918 influenza and COVID-19 pandemics, and (2) a synchronic analysis of COVID-19 coverage in the NYT and the more conservative New York Post (NYP). The aim is to understand how representations of expertise vary across time and media contexts, and whether political orientation influences expert portrayal.

The study employs methods from CADS (Partington, Duguid & Taylor 2013) on a corpus of 158 NYT articles (~148,000 words) and 80 NYP articles (~48,000 words). The analysis is conducted making use of word lists, concordances and a more detailed discourse analysis with Atlas.ti.

Findings indicate that during the 1918 pandemic, expert commentary in the NYT was typically mediated through officials, especially NYC Health Commissioner Dr. Copeland, rather than quoted directly. In contrast, during COVID-19, experts were frequently cited directly and were sometimes positioned as external commentators on political decisions. Furthermore, while the NYT drew on a wide range of expert voices, the NYP leaned more heavily on individual researchers and their personal stories.

This comparative analysis sheds light on the evolving role of experts in media narratives and the impact of political orientation on their representation during crises.

List of references

Partington, Alan, Alison Duguid, and Charlotte Taylor. 2013. Patterns and Meanings in Discourse: Theory and Practice in Corpus-Assisted Discourse Studies (CADS) [Studies in Corpus Linguistics 55]. John Benjamins.

Advancing corpus stylistics in Brazilian literature: A case study of Erico Verissimo's O Continente

Rozane Rodrigues Rebechi, Geovani Henrique Santos Souza

Universidade Federal do Rio Grande do Sul

Corpus stylistics, a field that applies corpus linguistics methods and tools to analyze literary style (McIntyre & Walker, 2019), remains underexplored in Brazilian Portuguese literary studies, although it is consolidated for identifying textual and literary stylistic patterns in works written in English (Ibrahim, 2022; McIntery & Walker, 2022, among others). A challenge to its broader adoption in Brazilian research is the limited capability of computational tools to process Portuguese texts, especially in the automatic annotation of semantic fields for corpus analysis. This study aims to test the Portuguese interface of the semantic annotator PyMUSAS, integrated into WMatrix (Rayson, 2009), by analyzing the literary style of Erico Verissimo's regionalist novel O Continente (Verissimo, 2013 [1949]). Comparing Verissimo's work, set in the southern State of Rio Grande do Sul, with an anthology by his contemporary northeastern author, Jorge Amado, the tool identified the most representative semantic fields of Verissimo's novel. These stylistic patterns were subsequently analyzed manually. The results corroborate conclusions from traditional literary criticism while revealing previously overlooked aspects of Verissimo's style. For instance, the prominent semantic domain 'Warfare, defence and the army; weapons' aligns with Bisol and Porto's (2025) discussion on the normalization of violence in the novel. Meanwhile, domains such as Anatomy and Physiology reveal a notable emphasis on elements like 'eyes,' 'head,' and 'hand,' which have received minimal attention in traditional stylistic studies. However, the findings also highlight limitations. Approximately 13.41% of the corpus was assigned to unmatched or incorrect semantic domains, likely due to gaps in the annotated Brazilian Portuguese lexis. These inaccuracies risk compromising the comprehensiveness of semiautomatic stylistic analyses. Improving automatic annotation accuracy is therefore essential for advancing corpus stylistics and informing its application in Brazilian literature studies.

List of references

Bisol, L. V.; Porto, L. T. Violência e memória: uma leitura do romance O continente, de Erico Verissimo. Navegações, [S. l.], v. 8, n. 2, p. 146–155, 2016.

Ibrahim, W. M. A. Utilizing corpus stylistics to facilitate literary analysis: An assessment of the effectiveness of semantic domains in identifying major literary themes in a selection of Charles Dickens novels. AJELP: Asian Journal of English Language and Pedagogy, 10(1), p. 114-138, 2022.

McIntyre, D.; Walker, B. Using corpus linguistics to explore the language of poetry: a stylometric approach to Yeats' poems. In: A. O'Keeffe; M. J. McCarthy (org) The Routledge Handbook of Corpus Linguistics. 2nd. ed. Routledge, 2022. pp. 499-516.

McIntyre, D.; Walker, B. Corpus Stylistics: Theory and Practice. Edinburgh: Edinburgh University Press, 2019.

Rayson, P. Wmatrix: a web-based corpus processing environment. Computing Department, Lancaster University, 2009.

Verissimo, E. O tempo e o vento parte I : o Continente 1 / o Continente 2. 4ª ed. São Paulo: Companhia das Letras, 2013 [1949].

Constructing Vulvar pain and Vulvodynia: A corpus-assisted study of pain discourses in scientific literature

Giorgia Riboni, **Angela Zottola**

University of Turin

Mentions of vulvar pain have been documented for over 100 years, but until the 1980s this condition was neither pathologized nor medicalized. Vulvodynia, a condition whose diagnosis is based on the presence of chronic vulvar pain has only recently been pathologized and has not yet been properly categorized (Bornstein et al. 2016). In line with theories on the collective understanding of illnesses (Conrad and Schneider, 1980) which consider discursive constructions as a tool for shaping societal perceptions of legitimacy and credibility regarding health conditions, this paper sees the invisibilization of vulvar pain and vulvodynia as something that results from the broader issue of devaluing women and their experiences.

This study investigates the linguistic and discursive resources utilised within scientific discourse to construct, frame and (de)legitimize vulvodynia within scholarly literature. To this end, SketchEngine was used to analyse a dataset of 141 academic papers published in English from 1998 to 2023. The investigation, based within corpus-assisted discourse studies (Gillings et al. 2023), was carried out by a synergical combination of quantitative and qualitative methods (Baker et al. 2008), starting with the identification of the most frequent lemmas (and the most salient topics; cf. van Dijk 2001: 102) and the examination of their contexts of occurrence (namely concordance and collocations). This analysis revealed a noticeable framing shift within the medical community (from a predominantly psychologized understanding of vulvar pain to a more biomedical one) and the pivotal role of patients' narratives in shaping the conceptualization of vulvodynia.

This study contributes to a broader understanding of how language constructs the reality of invisible pathologies that may ultimately lead to discriminatory practices; its findings have significant implications for improving medical communication and enhancing the quality of care for women.

List of references

Baker, Paul, Gabrielatos, Costas, KhosraviNik, Maijd, Krzyzanowski, Michal, McEnery, Tony & Ruth Wodak. 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 & Society 19(3): 273–306.

Bornstein, Jacob et al. 2016. 2015 ISSVD, ISSWSH and IPPS consensus terminology and classification of persistent vulvar pain and vulvodynia. Obstetrics & Gynecology 127, 745–751.

Conrad, Peter & Joseph W. Schneider. 1980. Deviance and Medicalization. From Badness to Sickness. Temple University Press.

Gillings, Mathew, Mautner, Gerlinde, and Paul Baker. 2023. Corpus-Assisted Discourse Studies. Cambridge University Press.

Van Dijk, Teun A. 2001. Multidisciplinary CDA: A Plea for Diversity. In Wodak, Ruth & Micheal Meyer (eds.) Medical of Critical Discourse Analysis. Sage: 95-120.

Analysing branching choices in videogame dialogue using the video game dialogue corpus

Sean G. Roberts[1], **Stephanie Rennick**[2]

[1]Cardiff University; [2]University of Stirling

Background

The Video Game Dialogue Corpus is an open-source corpus of scripted character dialogue from videogames (Rennick & Roberts, 2024). It represents branching dialogue choices, a unique feature of videogames that requires new analysis methods. For example, Rennick et al., (2023) used graph-walking algorithms to reveal gender biases in branching choices. Here, we present "Player Choice Adjusted Frequency" (PCAF): the expected frequency of a word (word count weighted by the probability of a player observing the word), divided by the total frequency of the word in all branches. This measures word frequency adjusted for branching structure. For example, if a word appears twice, but in two parallel top-level branches, then its expected frequency is 1 and PCAF = 50%. Since players frequently avoid 'evil' choices (Rennick & Roberts, forthcoming), we hypothesised that words with low PCAF scores would have low valency (i.e. negative affect).

Methods

Expected probability might be calculated by treating dialogue trees as Markov chains and calculating stable distributions. However, dialogues are rarely ergotic, and player choices depend on prior choices. Therefore, we created a graph-walking algorithm which assigns probabilities to branches, assuming players make random choices but do not re-visit branches. Warriner et al. (2013) provided word valence ratings.

Results

PCAF was calculated for 4.6 million words (93,541 types) from 25 games (mean PCAF = 45%). As predicted, there was a small but significant positive correlation between PCAF and valence ($r = 0.08$, $p < 0.001$). Non-linear modelling suggested the effect was mainly for extreme valence: high-valence words had above-average PCAF (e.g. "happiness", 60.3%; "joy",54.1%; "fun", 47.7%) but very low-valence words had low PCAF ("murder", 31.7%; "die", 32.9%; "kill", 25.9%). This suggests there are important patterns in the structure of dialogue choices. We continue to develop tools for exploring videogame dialogue, including an interactive tree visualiser (https://correlation-machine.com/VideoGameDialogueCorpus/BG3Explorer/).

List of references

Rennick, S., Clinton, M., Ioannidou, E., Oh, L., Clooney, C., E.T., Healy, E., & Roberts, S. G. (2023). Gender bias in video game dialogue. Royal Society Open Science, 10(5), 221095.

Rennick, S., & Roberts, S. G. (2024). The video game dialogue corpus. Corpora, 19(1), 93-106.

Rennick, S. & Roberts, S. G. (forthcoming). Optionality in video game dialogue. Game Studies.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. Behavior research methods, 45, 1191-1207.

Enhancing retrieval-augmented generation: A hybrid approach using vector embeddings and collocation analysis

Andressa Rodrigues Gomide[1], Luciana Dias de Macedo[2], Cornelia Plag[1], Maria Conceição Carapinha Rodrigues[1], Susana Antunes Ferreira[1], Natalia Sarnowska[1]

[1]Universidade de Coimbra; [2]Universidade Federal de Minas Gerais

Recent advancements in retrieval-augmented generation (RAG) have demonstrated the potential of large language models to access external databases for enhanced performance in natural language processing tasks. However, RAG's reliance on a single retrieval mechanism, typically based on dense vector embeddings, limits its ability to capture contextual nuances and linguistic patterns. This study explores how a hybrid retrieval approach combining (a) vector database searches with embeddings and (b) relational database searches for textual patterns can improve query accuracy and result relevance.

The data consisted of a corpus of 100 academic articles across multiple disciplines. Two databases were constructed: a vector database containing embeddings of the articles and a relational database storing the top collocations and their frequencies within each text. The list of top collocations was generated by first creating a lemma frequency list for each text. For each top lemma, collocates were identified using a Z-score measure. To search the relational database, the query was lemmatized, and the RapidFuzz Python library was employed to identify the most relevant texts based on collocation matches. Query accuracy was tested under three conditions: (1) vector database only, (2) relational database only, and (3) a hybrid approach where varying weights were assigned to the vector and relational databases.

The hybrid approach consistently outperformed single-database methods. Assigning higher weights to the vector database while incorporating collocation data from the relational database enabled more precise retrieval of contextually relevant information. These findings suggest that embedding-based retrieval can benefit from the complementary linguistic insights offered by collocation analysis. The results highlight the potential value of integrating diverse retrieval strategies in corpus linguistics and text mining, as combining semantic similarity with explicit linguistic patterns addresses gaps in existing RAG methodologies. Future research could extend this model to larger corpora and additional fields to further validate its applicability.

Mapping linguistic variation: A corpus-driven approach to forensic authorship profiling

Dana Roemling

University of Birmingham

Forensic authorship profiling is the analysis of texts to infer social characteristics of unknown authors - such as age, gender, or first language influence - by drawing on sociolinguistic work (Nini, 2018). However, inferring the regional background of an author has received limited attention, despite notable successes, such as identifying the regionalism “devil strip” in a ransom note to solve a high-profile case (Shuy, 2001). This study builds on such precedents by turning to corpus linguistics to analyse a large-scale dataset, providing new insights into authorship profiling through the lens of regional language use.

In this talk I present a corpus-driven approach to geolinguistic profiling, moving beyond traditional reliance on regionalisms identified by an analyst through dictionaries or dialect atlases. Using a corpus of 21 million geolocated German social media posts from the platform Jodel (Hovy & Purschke, 2018), I examine how regional linguistic variation can be systematically analysed at scale. I begin by evaluating the corpus for its regional distribution of linguistic features, addressing challenges of data sparsity and urban bias often encountered in geolocated datasets. By applying ordinary kriging (Wackernagel, 2003), a geospatial statistical method, I interpolate linguistic data for unobserved areas, enhancing spatial coverage and resolution. Furthermore, I draw on the corpus to develop an algorithm for dialect region prediction, showcasing how corpus-driven insights can inform forensic analyses and improve explainability within the legal context.

The findings of this study highlight the value of corpus linguistic methods in enhancing forensic authorship profiling. Not only does employing a large, geolocated corpus enable more accurate and scalable dialect region prediction, but it also provides a reference tool for qualitative analysis. By integrating corpus-driven approaches with forensic authorship profiling, this research supports the development of automated tools, further advancing the field of forensic linguistics.

List of references

Hovy, D., & Purschke, C. (2018). Capturing Regional Variation with Distributed Place Representations and Geographic Retrofitting. Proceedings of the 2018 Conference on EMNLP, 4383–4394. https://doi.org/10.18653/v1/D18-1469

Nini, A. (2018). Developing forensic authorship profiling. Language and Law / Linguagem e Direito, 5(2), 38–58.

Shuy, R. W. (2001). DARE’s role in linguistic profiling. DARE Newsletter, 4(3), 1–5.

Wackernagel, H. (2003). Multivariate Geostatistics: An Introduction with Applications. Springer.

How can usage-based SLA findings be translated into pedagogical practice? Examples from learner corpus research

Ute Römer-Barron

Georgia State University

Usage-based approaches to language that take language (in) use and its impact on language acquisition seriously have become increasingly popular and influential in applied linguistics (e.g., Cadierno & Eskildsen, 2015; Ellis & Wulff, 2020; Tyler et al., 2018). We have numerous insights from learner corpus research in the usage-based tradition on how language input affects second language (L2) acquisition. For instance, we know from this type of research that L2 learners are sensitive to frequency distributions in the input they receive (Ellis et al., 2016), that higher input frequencies facilitate stronger entrenchment of language features (Ellis, 2019), and that input enhancements may facilitate the fast learning of new constructions (Goldberg & Casenhiser, 2008). However, we know little about what exactly these insights could mean for the practice of L2 instruction and how pedagogical materials could be adjusted to better reflect usage-based acquisition principles.

The goal of this paper is to discuss how findings from usage-based SLA research could be translated into pedagogical practice. The paper will start with a brief overview of the main determinants of language learning from a usage-based perspective and will then summarize findings from recent learner-corpus studies on second language development (with a focus on verb constructions; e.g., Author, 2019, 2024), before presenting examples of pedagogical interventions that are inspired by these findings. The paper will provide specific recommendations for second language education and propose sample materials for the teaching of constructions and frames. It will also evaluate the effectiveness of pedagogical uses of corpora, notably data-driven learning (DDL; Johns, 1991), from a usage-based SLA perspective and list tasks for learner corpus researchers who wish to support language teachers and their students.

List of references

References
Author (2019)
Author (2024)
Cadierno, T., & Eskildsen, S. W. (Eds.). (2015). Usage-based Perspectives on Second Language Learning. Berlin: De Gruyter.
Ellis, N. C. (2019). Essentials of a theory of language cognition. The Modern Language Journal, 103, 39-60.
Ellis, N. C., Römer, U., & O'Donnell, M. B. (2016). Usage-based Approaches to Language Acquisition and Processing: Cognitive and Corpus Investigations of Construction Grammar. Malden, MA: Wiley.
Ellis, N. C., & Wulff, S. (2020). Usage-based approaches to L2 acquisition. In B. VanPatten, G. Keating, & S. Wulff (Eds.), Theories in Second Language Acquisition: An Introduction (pp. 63-82). London: Routledge.
Goldberg, A. E., & Casenhiser, D. (2008). Construction learning and second language acquisition. P. Robinson & N. C. Ellis (Eds.), Handbook of Cognitive Linguistics and Second Language Acquisition (pp. 197-215). London: Routledge.
Johns, T. F. (1991). Should you be persuaded—Two samples of data-driven learning materials. In T. F. Johns & P. King (Eds.), Classroom Concordancing. ELR Journal, 4, 1-16.
Tyler, A. E., L. Ortega, M. Uno, & H. I. Park (Eds.). (2018). Usage-Inspired L2 Instruction: Researched Pedagogy. Amsterdam: Benjamins.

Formulaic language and the development of grammatical knowledge: The case of caused-motion constructions

Rodrigo Garcia Rosa

University of Sao Paulo

The role of formulaic language in real-world communication is widely acknowledged in contemporary linguistic theory (Ellis, 2008, 2013; Wray, 2002). However, it wasn't until recently that cognitive models of language began to recognize its significance. This may stem from the long-standing dominance of modular views of language, which treated syntax and the lexicon as separate, independent systems. In contrast, recent cognitive linguistic approaches (Goldberg, 1995, 2006; Langacker, 2013) reject this rigid division, arguing that language lacks clear-cut boundaries (Evans, 2012; Lakoff, 1991). From a cognitive constructionist perspective, syntax and the lexicon are seen as a continuum of conventional *symbolic units* or *constructions*, varying in complexity, specificity, and schematicity. This framework accommodates both general, schematic argument structures and more fixed, conventional formulaic expressions that instantiate these structures. The phraseologisms in (1) and (2) below serve as specific, lexicalized examples of the broader schematic caused-motion construction, X CAUSES Y TO MOVE.

1. Conservatives have been more effective in *getting their ideas across to* the public. [COCA/1994/NEWS]
2. He knows how vital it is to *get the word out* that prostate cancer is the second most deadly cancer in men. [COCA/1992/MAG]

This talk explores the key grammatical features of English constructions commonly categorized as complex transitive constructions (Quirk et al., 1985), causative resultatives (Goldberg & Jackendoff, 2004), and caused-motion constructions (Goldberg, 1995). Drawing on empirical corpus-based research (Hampe, 2010), it argues that lower-level phraseological constructions, like *"talk some sense into"*, are essential in motivating the entrenchment and usage of more schematic constructions, such as *"Frank sneezed the foam off the cappuccino"*. To empirically support this argument, we analyze 1,284 utterances from COCA, identifying 12 fixed expressions and 9 statistically significant formulaic units. Finally, we consider the implications of this relationship between grammar and phraseology for understanding schematic structures like the caused-motion construction.

List of references

Ellis, N. 2008. "Phraseology: The Periphery and the Heart of Language". Phraseology in Language Learning and Teaching. Ed. F. Meunier and S. Granger. Amsterdam: John Benjamins. pp.01-13.

Ellis, N. 2013. "Construction Grammar and Second Language Acquisition". The Oxford Handbook of Construction Grammar. (Ed.) T. Trousdale, G. Hoffmann. New York: Oxford University Press, p. 15-31.

Evans, V. 2012. "Cognitive linguistics". Wiley Interdisciplinary Reviews: Cognitive Science, 3, p. 129.

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

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

Goldberg, A. E. and Jackendoff, R. 2004. "The resultative as a family of constructions". Language, 80, p. 532–68.

Hampe, B. 2010. "Metaphor, constructional ambiguity and the causative resultatives". Windows to the mind: Metaphor, metonymy and conceptual blending. Eds. S. Handl and H. J. Schmid Berlin & New York: Mouton de Gruyter, p. 185–215.

Lakoff, G. 1991. "Cognitive versus generative linguistics: How commitments influence results". Language & Communication, 11(1/2), pp.53-62.

Langacker, R. 2013. Essentials of Cognitive Grammar. Oxford: Oxford University Press.

The contextualisation of corpora through social media engagement metrics: Building a corpus of discourse around dementia on Twitter

Chris Sanderson

Lancaster University

Social media is an increasingly influential media – with 47% of 18-24 year olds saying they use it as a source of news (Newman et al., 2024). Social media also offers a vast wealth of naturally-occurring discourse - piquing the interest of corpus linguists (Rüdiger & Dayter, 2020). However, this vast amount of discourse often necessitates some kind of down-sampling.

A range of down-sampling methods can be identified from recent papers on social media discourse, such as selecting accounts with the most followers (Öneren-Özbek, 2024), collecting the posts which first appear when searching a particular hashtag or topic (Aboh, 2024) and taking all posts which include particular keyword(s) (Coltman-Patel et al., 2022). However, Kwak et al (2010)'s study suggests that follower count is not linked to engagement, the algorithm decides what to show a user based on their individual behaviour (Twitter, 2023), and the importance of contextual aspects such as text production, interaction and participation is increasingly emphasised (Herring, 2013; KhosraviNik, 2017).

This study aims to provoke a critical discussion of these methods, and explore their potential implications for linguistic findings. This paper does this through a close investigation of the metadata of 18.9 million tweets from 2013-2022 which contain either *dementia* or *Alzheimer's*. The results demonstrate the value in incorporating engagement metrics – enabling comparison of what is **on** Twitter against what is **seen** on Twitter. It will also challenge intuitive assumptions about the statistical relationship between metrics such as likes, retweets, replies and follower count.

To further explore these findings, I will critically engage with the impact of the platform's algorithm and the different intentions behind engagement – such as the personalized 'reply' against the more public 'retweet' as explored by Segev (2023).

List of references

References

Aboh, S. C. (2024). 'It will never be well with SARS': A discourse analytic study of the #EndSARS protests on social media. Discourse & Society, 35(2), 153–173. https://doi.org/10.1177/09579265231200994

Coltman-Patel, T., Dance, W., Demjén, Z., Gatherer, D., Hardaker, C., & Semino, E. (2022). 'Am I being unreasonable to vaccinate my kids against my ex's wishes?' – A corpus linguistic exploration of conflict in vaccination discussions on Mumsnet Talk's AIBU forum. Discourse, Context & Media, 48, 100624. https://doi.org/10.1016/j.dcm.2022.100624

Herring, S. (2013). Discourse in Web 2.0: Familiar, Reconfigured, and Emergent. In D. Tannen & A. M. Trester, Discourse 2.0: Language and new media (pp. 1–25). Georgetown University Press.

KhosraviNik, M. (2017). Social media critical discourse studies (SM-CDS). In J. Flowerdew & J. E. Richardson (Eds.), The Routledge Handbook of Critical Discourse Studies (1st ed., pp. 582–596). Routledge. https://doi.org/10.4324/9781315739342-40

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media? Proceedings of the 19th International Conference on World Wide Web, 591–600. https://doi.org/10.1145/1772690.1772751

Newman, N., Fletcher, R., Robertson, C. T., Ross Arguedas, A., & Nielsen, R. K. (2024). Reuters Institute Digital News Report 2024. Reuters Institute for the Study of Journalism. https://doi.org/10.60625/RISJ-VY6N-4V57

Öneren-Özbek, M. (2024). Antisemitism in contemporary Türkiye: Discourses on Turkish Jews on Twitter. Discourse & Society, 35(1), 116–136. https://doi.org/10.1177/09579265231195461

Rüdiger, S., & Dayter, D. (Eds.). (2020). Corpus Approaches to Social Media (Vol. 98). John Benjamins Publishing Company. https://doi.org/10.1075/scl.98

Segev, E. (2023). Sharing Feelings and User Engagement on Twitter: It's All About Me and You. Social Media + Society, 9(2), 20563051231183430. https://doi.org/10.1177/20563051231183430

Twitter. (2023). Twitter's Recommendation Algorithm. https://blog.twitter.com/engineering/en_us/topics/open-source/2023/twitter-recommendation-algorithm

Uncovering diachronic trends in social media corpora: A usage-fluctuation-analysis of discourse around dementia on Twitter

Chris Sanderson

Lancaster University

Social media is often characterised by transient trends, and studies sometimes look at short time periods or events (Aboh, 2024; Russo, 2023), others condense multiple years (Coltman-Patel et al., 2022; Lorenzo-Dus & Nouri, 2021). It is unclear to what extent these trends and events influence ongoing discourse on social media, or whether the discourse is robust to change. If the discourse is in flux, this might complicate looking at social media discourse from a longitudinal perspective.

This study applies a Usage Fluctuation Analysis (UFA) to a corpus of 18.9 million tweets from 2013-2022 which contain either *dementia* or *Alzheimer's*. UFA is well suited to this question as it enables a look into how the collocates around a node word have changed over time (McEnery et al., 2019). However, implementing UFA requires navigating a number of 'moving parts' such as window size, window interval, collocate parameters and graph smoothing, which can present a "steep learning curve" (Gillings & Dayrell, 2023). This paper outlines the methodological decisions made in addressing these challenges, and highlights their analytical implications.

The dispersion of collocates across different parameter adjustments was used to guide parameter selection, and enabled the analysis to involve sufficient collocates to be meaningful, whilst mitigating the uneven distribution of textual evidence across different temporal windows. The study also emphasizes the triangulation of techniques such as consistent collocates and concordancing to provide a comprehensive interpretation of the usage changes demonstrated by the UFA graph.

The resultant analysis demonstrates the potential value of longitudinal social media analysis and the use of UFA in uncovering diachronic trends. It also highlights how filtering tweets by engagement yields a more nuanced understanding of how transient trends impact social media discourse over time.

List of references

References

Aboh, S. C. (2024). 'It will never be well with SARS': A discourse analytic study of the #EndSARS protests on social media. Discourse & Society, 35(2), 153–173. https://doi.org/10.1177/09579265231200994

Coltman-Patel, T., Dance, W., Demjén, Z., Gatherer, D., Hardaker, C., & Semino, E. (2022). 'Am I being unreasonable to vaccinate my kids against my ex's wishes?' – A corpus linguistic exploration of conflict in vaccination discussions on Mumsnet Talk's AIBU forum. Discourse, Context & Media, 48, 100624. https://doi.org/10.1016/j.dcm.2022.100624

Gillings, M., & Dayrell, C. (2023). Climate change in the UK press: Examining discourse fluctuation over time. Applied Linguistics, 45(1), 111–133. https://doi.org/10.1093/applin/amad007

Lorenzo-Dus, N., & Nouri, L. (2021). The discourse of the US alt-right online – a case study of the Traditionalist Worker Party blog. Critical Discourse Studies, 18(4), 410–428. https://doi.org/10.1080/17405904.2019.1708763

McEnery, T., Brezina, V., & Baker, H. (2019). Usage Fluctuation Analysis: A new way of analysing shifts in historical discourse. International Journal of Corpus Linguistics, 413–444. https://doi.org/10.1075/ijcl.18096.mce

Russo, K. (2023). Fear Appeals, Migration and Sinophobia in COVID-19 News and Twitter Discourse: A Corpus-based Critical Analysis. Critical Approaches to Discourse Analysis across Disciplines, 14(2), 21–40.

The role of corpus size and part-of-speech in enhancing collocation learning through DDL

Yoshiho Satake

Aoyama Gakuin University

This study investigates the effects of corpus size, part-of-speech focus, and target word difficulty on the efficacy of data-driven learning (DDL) among Japanese university students. It explores how corpus size influences the availability and effectiveness of collocations, whether the grammatical category of collocations (adjectives and verbs) affects learning outcomes, and whether target word difficulty impacts DDL effectiveness. Participants were 19 Japanese first-year university students (CEFR B1-B2) using a large-scale corpus (enTenTen) and 21 using a small-scale corpus (Brown Corpus) in a weekly Reading & Writing course. Both groups received a 20-minute instructional session on corpus usage before the treatment, which involved four weeks of DDL activities. Each session focused on 2-4 nouns from a university-designated textbook, ranging from CEFR A2 to C2. Students searched for common collocations (preceding adjectives, preceding verbs, and following verbs) using the assigned corpus, completed worksheets, and composed sentences using the collocations, followed by reflections. Pre- and post-tests measured learning outcomes, and surveys provided qualitative insights. Results showed that while the small-scale corpus sometimes lacked sufficient collocation data, there was no significant difference between groups in post-test performance regarding memory and productive use of collocations. Both groups identified suitable adjectives and verbs during sessions, but the small corpus group faced limitations due to reduced collocation availability. Additionally, participants often recalled high-frequency, lower-difficulty adjectives in post-tests, while verb collocations seemed more beneficial for learning novel patterns. Target word difficulty influenced pre-test results but not post-test results, suggesting that DDL can facilitate collocation learning regardless of word complexity. These findings contribute to research on effective DDL practices by highlighting the balance between corpus size and learning outcomes and the role of part-of-speech and word difficulty in collocation acquisition.

Investigating discipline-specific academics' expectations of undergraduate writing

Analisa Scerri

University of Malta

Academic writing is a social and linguistic passport that offers students access to and membership of shared professional practices and disciplinary communities. As such, it is crucial for students to obtain both disciplinary knowledge as well as skills in effectively communicating that knowledge. For the latter, student writing is worthy of attention given its role as the "predominant medium through which students are assessed during their academic careers." (Murray and Sharpling, 2019, p. 489). However, while academic writing is key in shaping and evaluating students' academic success in Higher Education, what is expected of good disciplinary academic writing is not always easy to determine (Lillis & Turner, 2011). Considering this lack of clarity surrounding writing expectations, this paper analyses discipline-specific academics' expectations of undergraduate writing through interview-based research with academics at the University of Malta – academics who serve as the primary readers and assessors of students' disciplinary writing (Hyland, 2018).

Adopting a corpus-based approach informed by critical grounded theory, this study outlines the parameters of high-quality writing through two research aims. The first aim involves examining the expectations of these academics, which they articulate through interviews and evaluations of students' work. The second aim focuses on the discursive construction of these expectations, in relation to students' undergraduate writing. Thus the analysis combines content analysis with corpus-assisted discourse analysis and sheds light on an occluded facet of academic practice. While reporting on the analysis, the presentation also reflects on the challenges apparent in conducting this study which includes a focus on the pilot analysis, the original methodological framework developed, and the analytical tools explored in the process. Significant changes have been introduced as a result of this pilot which was conducted to refine the methodology and research questions of a Ph.D. study in an area that has received limited attention.

List of references

Hyland, K. (2018). The Essential Hyland. Bloomsbury Academic.

Lillis T., & Turner J., (2011). Student Writing in Higher Education: Contemporary confusion, traditional concerns. Teaching in Higher Education, 6:1, 57-68.

Murray, N., and Sharpling, G. (2019) What traits do academic value in student writing? Insights from a psychometric approach. 44:3, 489-500.

Hidden in plaintext: Transforming archival metadata into corpora

Hanna Schmück[1,2], **Marc Alexander**[1]

[1]University of Glasgow; [2]University of Augsburg

This paper explores the potential of utilising archival materials as corpora, emphasizing the cultural value of language as a heritage object. Despite the wealth of community archives across the UK, these resources remain underappreciated and underutilised in corpus linguistic research. Archives encompass different linguistic varieties and encompass a wide range of subjective and politically charged viewpoints (Fitzgerald, 2022), not to mention having great potential for corpus researchers as their data often contain a range of different genres and modes. Inspired by Pagenstecher and Pfänder (2017), this paper promotes cooperation between archivists and linguists to create archival corpora. Unlike previous work on transforming historical oral testimonies into spoken corpora (see Clary-Lemon, 2010; Fitzgerald, 2022), we focus on constructing corpora from secondary textual information.

Two core research questions guide this study: How can we transform archival data into corpora? What can archival corpora reveal about communities of practice? Our case study involves transforming the archival metadata from the People's Collection Wales (PCW) into two corpora: one containing over 7.7 million tokens of English descriptions from 153,000 metadata files, and the other containing the Welsh descriptions amounting to 6.8 million tokens from the same number of metadata files. This metadata includes descriptions of community-generated digital content, such as scans of photographs, transcriptions of letters, interviews, and videos, often contributed by volunteers. Our paper details the practical steps of identifying, collecting, and cleaning archival data for corpus research, illustrating the unique insights that can be uncovered through such archival materials compared to existing corpora. We then employ collocation and keyness analyses to explore cultural themes within either corpus and highlight insights into local communities of practice and cultural differences. This research outlines a pathway for corpus linguistics that aims for a more community-based, culturally aware, and interdisciplinary approach to language.

List of references

References

Clary-Lemon, J. (2010). 'We're not ethnic, we're Irish!': Oral histories and the discursive construction of immigrant identity. Discourse & Society, 21(1), 5–25. https://doi.org/10.1177/0957926509345066

Fitzgerald, C. (2022). Investigating a Corpus of Historical Oral Testimonies: The Linguistic Construction of Certainty (1st ed.). Routledge. https://doi.org/10.4324/b22799

Pagenstecher, C., & Pfänder, S. (2017). Hidden dialogues: Towards an interactional understanding of oral History interviews. In Erich Kasten, Katja Roller, & Joshua Wilbur, Oral History Meets Linguistics (pp. 185–207).

Roller, K. (2015). Towards the 'oral' in oral history: Using historical narratives in linguistics. Oral History, 43(1), 73–84.

Teaching corpus linguistics in context(s): Reflections and future directions

Maya Sfeir

American University of Beirut

Even with the expansion of corpus linguistics as a discipline—as evident in the increasing number of relatively new journals dedicated to the field, such as *Research in Corpus Linguistics* and *Applied Corpus Linguistics*—little attention has been given to the teaching of corpus linguistics (thereafter CL) as a subject. With a few exceptions such as Renouf's (1997) chapter, the discussion surrounding the teaching of CL has been largely overlooked since Fligelstone (1993) advanced his three-tiered model for teaching about corpora. Reports on the practical dimension of teaching CL without adequate theoretical grounding (see Stoykova, 2014) add another layer to the challenge. Two contradictions thus arise. While CL as a field brings together a vibrant, active, open, and supportive—even generous—community of scholars, discussions surrounding the teaching of CL are largely neglected. Second, whereas data-driven learning as a sub-field within CL calls for the integration of corpus methods and data in language teaching and learning, the teaching of CL itself remains unexplored. In my presentation, I build on (limited) existing discussions of teaching CL to inquire: what strategies can amplify the conversation on teaching CL in (local) context(s)? I draw on my experience—and relevant data in the form of syllabi, assignments, and student reflections—designing, teaching, and revising corpus linguistics courses and modules at a public and a private higher education institution in the Middle East. In light of my experience and the scarce literature on the teaching of CL, I emphasize three (future) foundations for teaching CL in (local) context(s): communities (including students), resources (data/corpora and tools), and innovations (to overcome contextual constraints). My presentation ultimately seeks to emphasize the "*how*" and "*to whom"* we teach CL that Fligelstone (1993, p. 98) referred to—with the ultimate aim of positioning linguistic and non-linguistic majors as "language researchers" (Cheng et al., 2003, p. 178).

List of references

Cheng, W., Warren, M., & Xun-Feng, X. (2003). The language learner as language researcher: Putting corpus linguistics on the timetable.System,31(2), 173–186.

Fligelstone, S. (1993) Some reflections on the question of teaching, from a corpus linguistics Perspective. ICAME Journal, 17, 97–110.

Renouf, A. (1997) Teaching corpus linguistics to teachers of English. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (Eds.), Teaching and language corpora (pp. 255–66). Longman.

Stoykova, V. (2014). Teaching corpus linguistics. Procedia-Social and Behavioral Sciences, 143, 437-441.

Typological differences, directionality, and cognitive load in manner translation: A corpus-based study of Chinese-English and English-Chinese consecutive interpreting

Lin Shen

University of Cambridge

The influence of inter-typological variations on the processing of manner information has been extensively examined in the domain of motion. Manner, however, extends to more other semantic domains, as demonstrated in the onomasiological approach to manner analysis. This study, based on this approach, analyzes the influence of directionality and cognitive load (measured by interpreting performance) on the transfer of manner under high cognitive demands, using bidirectional corpus data of consecutive interpreting between Chinese (an equipollently-framed language) and English (a satellite-framed language). The results indicate that 1) increased interpreting performance correlates with higher transfer rates of both manner adjuncts and verbs; 2) transfer rates for manner verbs are significantly higher when interpreting into English (CE) than into Chinese (EC), supporting previous findings on the salience of manner in English; 3) interpreting direction influences resistance to cognitive load, with manner adjuncts showing greater resistance in the EC direction due to Chinese's more flexible locus of manner information, while manner verbs exhibit better resistance in the CE direction, reflecting English's higher verbal codability of manner. These findings suggest the broader applicability of Talmy's typology to semantic domains beyond motion and to processing under high cognitive loads in different language combinations and interpreting modes.

List of references

Combe, C., & Stosic, D. (2024). Processing manner under high cognitive pressure: Evidence from French–English and English–French simultaneous interpreting. Language and Cognition, 1–29. https://doi.org/10.1017/langcog.2023.74

Lv, Q., & Liang, J. (2019). Is Consecutive Interpreting Easier than Simultaneous Interpreting? - A Corpus-Based Study of Lexical Simplification in Interpretation. Perspectives, 27(1), 91–106. https://doi.org/10.1080/0907676X.2018.1498531

Moratto, R., & Yang, Z. (2023). Probing the cognitive load of consecutive interpreters: A corpus-based study. Translation and Interpreting Studies. https://doi.org/10.1075/tis.22047.mor

Slobin, D. I. (2004). The many ways to search for a frog: Linguistic typology and the expression of motion events. In S. Sven & L. Verhoeven (Eds.), Relating events in narrative, vol. 2: Typological and contextual perspectives (pp. 219–257). Lawrence Erlbaum.

Stosic, D. (2020). Defining the Concept of Manner: An Attempt to Order Chaos. UniSa. Sistema Bibliotecario Di Ateneo. https://doi.org/10.14273/unisa-3440

Talmy, L. (2000). Toward a cognitive semantics. MIT Press.

Wen, Q., & Wang, J. (2008). Parallel Corpus of Chinese EFL Learners. Foreign Language Teaching and Research Press.

Simulating human artificiality: How well does AI generate EFL textbook texts?

Marilisa Shimazumi[2], Tony Berber Sardinha[1]

[1]Pontifical Catholic University of Sao Paulo; [2]Cultura Inglesa College

The use of AI-generated texts in educational materials is an emerging area of interest, particularly in crafting language teaching example texts and tasks. Specifically, this study investigates whether AI-generated texts can replicate the characteristics of English coursebook texts, which are considered artificial because they are designed to meet specific pedagogical goals rather than the communicative demands of real-world contexts (Carter, 1998; Gilmore, 2004). As such, it could be argued that generating EFL texts with AI involves artificially reproducing instructional artificiality. The research analyzed the English Language Teaching Textbook (ELTT) corpus, comprising 106,840 words from 500 texts across 19 registers, evenly split between B2 and C1 levels. The corpus was tagged with the Biber Tagger and post-processed with the Biber TagCount program. A Multi-Dimensional Analysis (Biber, 1988) identified five dimensions of variation: (1) persuasion, speaker engagement, and personal opinion vs. analysis and technical information; (2) expressive, interactive, speculative discourse with stance marking; (3) formal, detailed, and informative composition; (4) narrative and descriptive accounts; and (5) summarized abstract overviews. A comparison corpus (AI-ELTT) of 500 simulated textbook texts was created with ChatGPT-4, replicating the coursebook texts. The AI was prompted to match the proficiency level, register and topic of the target texts through careful prompt design. An additive MD Analysis (Berber Sardinha et al., 2019) of the AI-ELTT texts on the ELTT dimensions revealed significant differences. AI struggled to replicate persuasive and interactive language, often favoring technical and analytical expressions. It also lacked the stance marking and speculative discourse typical of human-authored texts. A Discriminant Functional Analysis revealed that the dimension scores allowed AI and human texts to be distinguished in over 80% of cases. The findings demonstrate that although AI can mimic some aspects of textbook language, it falls short of reproducing the exact kind of artificiality of EFL coursebook texts.

List of references

Berber Sardinha, T., Veirano Pinto, M., Mayer, C., Zuppardi, M. C., & Kauffmann, C. (2019). Adding registers to a previous Multi-Dimensional Analysis. In T. Berber Sardinha & M. Veirano Pinto (Eds.), Multi-Dimensional Analysis: Research Methods and Current Issues (pp. 165-188). London: Bloomsbury Academic.

Biber, D. (1988). Variation Across Speech and Writing. Cambridge: Cambridge University Press.

Carter, R. (1998). Orders of reality: CANCODE, communication, and culture. ELT Journal, 52, 43-56.

Gilmore, A. (2004). A comparison of textbook and authentic interactions. ELT Journal, 58(4), 363-374.

A two-dimensional keyness analysis of person marking constructions in Hungarian lyric poetry – theoretical possibilities and methodological challenges

Gábor Simon, Emese K. Molnár

Eötvös Loránd University

The paper aims to explore the significant patterns of person marking in three subcorpora of the Corpus of Hungarian Lyrical Poetry (Horváth–Simon–Tátrai 2022), namely canonical lyric poetry, contemporary lyric poetry, and song lyrics. The analysis is motivated by the ongoing corpus-based research in the Stylistic Research Group according to which person marking constructions constitute genre-specific patterns in lyrical discourses. The paper investigates the empirical relevance of this statement.

Adopting the two-dimensional keyness analysis proposed by Gries (2021), we compared all the lyrical texts with the data of the ELTE Novel Corpus (Bajzát–Szemes–Szlávich 2021) on the one hand. We contrasted on the other hand the individual lyrical subcopora with the rest of the corpus. As a keyness measurement, the Kullback-Leibler divergence is used to identify the key grammatical features of Hungarian lyric poetry in general and its various fields. The quantitative analyses are performed in RStudio.

According to the preliminary results, second-person singular forms proved to be key features of lyrical discourse in general. In contrast, third-person singular forms dominate contemporary poetry, and second-person singular expressions of person marking are prominent in song lyrics, however, with particular semantic roles. Our study, thus, takes the corpus linguistic investigation of song lyrics (see Werner 2023a, 2023b, Langenhorst–Frommherz–Meier-Vieracker 2023) one step further. Moreover, it demonstrates a new application of corpus linguistic tools in stylistic and literary research.

List of references

References

Bajzát, Tímea Borbála – Szemes, Botond – Szlávich, Eszter 2021. Az ELTE DH Regénykorpusz és lehetőségei. [The ELTE DH Novel Corpus and its applications.] Networkshop 2021: 63–72. DOI: 10.31915/NWS.2021.7

Gries, Stefan Th. 2021. A new approach to (key) keyword analysis: Using frequency, and now also dispersion. Research in Corpus Linguistics 9 [2]: 1-33.

Horváth, Péter – Simon, Gábor – Tátrai, Szilárd 2022. Annotation of person marking constructions in the Corpus of Hungarian Lyrical Poetry. Studia Linguistica Hungarica 34: 22–37.

Langenhorst, Jan – Frommherz, Yannick – Meier-Vieracker, Simon 2023. Keyness in song lyrics: Challenges of highly clumpy data. Journal for Language Technology and Computational Linguistics. Special Issue on Challenges in Computational Linguistics, Empirical Research & Multidisciplinary Potential of German Song Lyrics. 36 [1]: 21–38.

Werner, Valentin 2023a. “Guess who’s back, back again.” Stylistic development in Eminem’s lyrics. In Schubert, Christoph - Werner, Valentin (Eds.): Stylistic approaches to pop culture. London: Routledge, 176–204.

Werner, Valentine 2023b. English and German pop song lyrics: Towards a contrastive textology. Journal for Language Technology and Computational Linguistics. Special Issue on Challenges in Computational Linguistics, Empirical Research & Multidisciplinary Potential of German Song Lyrics. 36 [1]: 1–20.

Association rules mining: A method for pattern discovery on sign language corpus data

Robert Smith, Markus Hofmann

Technological University Dublin

Since work on the first digital sign language corpora began in 2004 (e.g., Johnston & Schembri, 2005; Leeson et al., 2006), researchers have been deploying standard corpus analysis techniques such as n-gram analysis (Wolfe et al., 2014), concordance analysis (Crasborn et al., 2014), and lexical frequency analysis (Smith & Hofmann, 2020). The application of data mining and exploratory statistical methods is a recent trend for the analysis of sign language data, with such methods reported in Ferrara et al. (2020, 2022), Sallandre et al. (2019) and Saunders (2022). Building on this new direction, this study exploits a data science approach known as Association Rules Mining (ARM) with the aim to empirically identify patterns between the physical movements of Non-manual Features (NMFs) (head, body, and face movement) and lexical units in Irish Sign Language.

The study follows the Knowledge Discovery on Data (KDD) methodology to pre-process and analyse data from the Signs of Ireland Corpus (Leeson et al., 2006) using the R package arules implementation of the Apriori algorithm (Hahsler, 2021). The principal challenge facing this method is the comparatively small sign language corpora available for training machine learning models.

In addition to outlining our method, this presentation will report patterns discovered between grammatical classes and various non-manual articulations. One such pattern discovery is the strong correlation between various NMFs and depicting verbs. Indeed, this study reports that the less lexicalised a sign is, the more likely it is to include NMFs. Findings from this work provide linguistic insights based on empirical evidence. In addition, the findings will inform future research on NMF treatment in sign language processing, while the proven method provides a template to deploy the ARM method on similar data in future studies.

List of references

Crasborn, O., Hulsbosch, M., Lampen, L., & Sloetjes, H. (2014). New multilayer concordance functions in ELAN and TROVA. Tilburg Gesture Research Meeting [TiGeR 2013].

Ferrara, L., Anible, B., Hodge, G., Jantunen, T., Leeson, L., Mesch, J., & Nilsson, A.-L. (2020). A cross-linguistic comparison of reference across different signed languages. Proceedings of the 14th High Desert Linguistics Society Virtual Conference (HDLS), Albuquerque, New Mexico.

Ferrara, L., Anible, B., Hodge, G., Jantunen, T., Leeson, L., Mesch, J., & Nilsson, A.-L. (2022). A cross-linguistic comparison of reference across five signed languages. Linguistic Typology.

Hahsler, M. (2021). arules package documentation: Mining Association Rules and Frequent Itemsets.

Johnston, T., & Schembri, A. (2005). The use of ELAN annotation software in the Auslan Archive/Corpus Project. Proceedings of the Ethnographic E-Research Annotation Conference, University of Melbourne, Victoria, Australia, 1516.

Leeson, L., Saeed, J., Byrne-Dunne, D., Macduff, A., & Leonard, C. (2006). Moving Heads and Moving Hands: Developing a Digital Corpus of Irish Sign Language. Information Technology and Telecommunications. Carlow, Ireland, 2526.

Sallandre, M.-A., Balvet, A., Besnard, G., & Garcia, B. (2019). Étude exploratoire de la fréquence des catégories linguistiques dans quatre genres discursifs en LSF. Lidil. Revue de Linguistique et de Didactique Des Langues, 60.

Saunders, D. (2022). An analysis of adverbial and stance-taking mouth actions within and outside enactments in LSQ. Proceedings of Sign CAFE 2: 2nd International Workshop on Cognitive and Functional Explorations in Sign Language Linguistics.

Smith, R. G., & Hofmann, M. (2020). A Lexical Frequency Analysis of Irish Sign Language. TEANGA, the Journal of the Irish Association for Applied Linguistics, 11, 1847.

Wolfe, R. J., McDonald, J., Berke, L., & Stumbo, M. (2014). Expanding n-gram analytics in ELAN and a case study for sign synthesis. LREC, 1880–1885.

Exploring discursive multidimensionality and multimodality on Twitter: Analyzing xenophobic representations targeting China during the COVID-19 pandemic

Cicero Soares da Silva

Sao Paulo Catholic University - PUCSP

In this paper, we look at malicious representations of China on Twitter occurring in the context of the COVID-19 pandemic. In order to capture these detrimental representations, we scraped a corpus of ca. 100K tweets in Brazilian Portuguese containing ten highly xenophobic hashtags, which were used by right-wing followers to discredit China and spread hatred. The multimodal method followed in this study consisted of a combination of Lexical Multidimensional Analysis (LMDA; Berber Sardinha & Fitzsimmons-Doolan, 2024) and Visual Multidimensional Analysis (VMDA; Berber Sardinha et al., 2023). The LMDA used lexical units to detect traces of discourses across the texts, whereas the VMDA applied computer vision AI techniques to annotate the images posted along with the twitter messages. Two sets of dimensions were obtained (i.e. verbal and a visual). Six verbal dimensions were identified, the first two being: (1) Pandemic manipulation vs Pro-president hashtags: this contrasts the alleged manipulation involving the protection of corrupt officials and misleading pandemic data with the use of hashtags to ridicule China, support presidential policies, and target political adversaries. (2) System and media rejection hashtags vs Anti-China normalization: This captures the use of hashtags to reject the political system and its alleged ties with China, versus efforts to normalize anti-Chinese sentiments under the guise of common sense and cultural distortion. In turn, five visual dimensions were determined, the first two being: (1) China scam denouncement vs. Brazil Elite Accusations: This captures the contrasting positions of denouncing scams from China and accusing Brazilian intellectuals of siding with Chinese interests; (2) Weak local government pandemic response vs. Brave pandemic leadership: This contrasts the views on local governments' responses to the pandemic, either as weak and closure-promoting or brave and against closures seen as benefiting Chinese interests. All the dimensions will be discussed and illustrated in the paper presentation.

List of references

Berber Sardinha, T., & Fitzsimmons-Doolan, S. (2024.). Lexical Multidimensional Analysis. Cambridge University Press.

Berber Sardinha, T. (2024). Exploring multimodal corpora in the classroom from a multidimensional perspective In P. Crosthwaite (Ed.), Corpora for Language Learning: Bridging the Research-Practice Divide (pp. 25-36). Abingdon: Routledge.

Raising the bar: An analysis of bar chart usage in corpus data visualization

Lukas Sönning

University of Bamberg

According to a recent review of statistical graphics in corpus-based work, the bar chart is the most widely used graph type (Sönning & Schützler 2023). The current study expands this survey to cover all research articles published in five corpus-linguistic journals up to and including the year 2024 (*International Journal of Corpus Linguistics*, *Corpus Linguistics and Linguistic Theory*, *Corpora*, *Research in Corpus Linguistics*, and the *International Journal of Learner Corpus Research*). Out of the 1,165 papers in the report pool, 631 (54%) include some form of data visualization. In this subset, bar charts were used in 441 studies (70%). This usage rate is remarkably stable across time, suggesting that bar charts are here to stay. This is to be applauded – after all, despite the bad press they have occasionally gotten in the literature (e.g. Tufte 2001, Robbins 2005, Cleveland 1994), their wide familiarity and simplicity give them a key advantage over competing forms (Kosslyn 1985). The current study takes a closer look at how bar charts are used in corpus-linguistic papers. We start with a critical review of the literature on design principles and recommendations for (this form of) data visualization and then evaluate the layout of the 996 bar charts in our survey. This exercise pursues two goals. The first is to identify aspects where there is room for improvement. Our analysis shows that this primarily concerns issues related to scaling and absolute figure size and the arrangement of categories in the chart. The second goal is to establish contexts where an alternative graph type may be preferable. These chiefly concern more complex and multifactorial data layouts, where (stacked and grouped) bar charts quickly become cluttered, which interferes with the perception of patterns in the data.

List of references

Cleveland, William S. 1994. The elements of graphing data. Summit: Hobart Press.

Kosslyn, Stephen M. 1985. Graphics and human information processing: A review of five books. Journal of the American Statistical Association 80. 499–512.

Robbins, Naomi B. 2005. Creating more effective graphs. Hoboken: Wiley.

Sönning, Lukas & Ole Schützler. 2023. Data visualization in corpus linguistics: Critical reflections and future directions. In Lukas Sönning & Ole Schützler (eds.), Data visualization in corpus linguistics: Critical reflections and future directions. Helsinki: VARIENG. https://urn.fi/URN:NBN:fi:varieng:series-22-0

Tufte, Edward R. 2001. The visual display of quantitative information. Cheshire: Graphics Press.

Case-control down-sampling in corpus-based research

Lukas Sönning

University of Bamberg

When the list of hits returned by a corpus query is too large for a linguistic analysis, researchers can rely on a number of down-sampling strategies to carefully down-size their data (see Sönning 2024). Variationist research may then rely on case-control down-sampling, where the selection of tokens is based on the observed value of the outcome variable. A study of alternating constructions, for instance, may decide to select the same number of tokens per variant. We report the results of a survey of down-sampling practices in corpus-based work (based on 5,234 linguistic research articles published between 2012 and 2023 across 17 journals). We observe that this design is used at a noticeable rate: 15 out of 23 corpus-based studies (65%) that down-sample a set of data that allows for a response-sensitive design made use of this strategy. In other disciplines, most notably the health sciences, a rich methodology has evolved for case-control studies (Cornfield 1951; Breslow & Day 1980; Schlesselman 1982; Keough & Cox 2014; Borgan et al. 2018). The purpose of the present paper is to sound out the potential for methodological transfer to (variationist) corpus linguistics, where relevant know-how seems to be largely lacking. It pursues three goals. The first is to translate the rather idiosyncratic (but firmly entrenched) terminology into the language of our field. The proposed mapping of technical terms provides a basis for methodological exchange. The second goal is to summarize established guidelines relating to study design and data analysis. Finally, we will take a closer look at the 15 case-control studies in our survey to consider how corpus data layouts differ from the research contexts for which this sampling design was originally developed. This will draw our attention to aspects that require elaboration if corpus linguistics is to develop its own case-control methodology.

List of references

Borgan, Ørnulf, Norman E. Breslow, Nilanjan Chatterjee, Mitchell H. Gail, Alastair Scott & Christopher J. Wild. 2018. Handbook of statistical methods for case-control studies. Boca Raton, FL: CRC Press.

Breslow, Norman E. & N.E. Day. 1980. Statistical methods in cancer research: Volume 1 – The analysis of case-control studies. Lyon: International Agency for Research on Cancer.

Cornfield, Jerome. 1951. A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast, and cervix. Journal of the National Cancer Institute 11, 1269-1275.

Keogh, Ruth H. & David R. Cox. 2014. Case-control studies. Cambridge: Cambridge University Press.

Schlesselman, James J. 1982. Case-control studies: Design, conduct, analysis. Oxford: Oxford University Press.

Sönning, Lukas. 2024. Down-sampling from hierarchically structured corpus data. International Journal of Corpus Linguistics 29(4), 507-533.

It's no wonder headteachers are leaving! Utilising corpus methods to illuminate the policy problem of Ofsted school inspections in England

Kathryn Anne Spicksley

University of Birmingham

In England, state-maintained schools are regularly inspected by the Office for Standards in Education, Children's Services and Skills. The outcome of each inspection is a public report, the consequences of which can have serious ramifications for schools judged to be underperforming. The tragic suicide of the Reading headteacher Ruth Perry in the wake of a 2022 Ofsted inspection raised public awareness of the high-stakes nature of Ofsted reports. However, despite the potential impact of these reports on schools, teachers and pupils, there is limited academic research undertaken on the discursive patterns evident in Ofsted reports utilising corpus methods.

This paper reports on a corpus-assisted critical discourse analysis conducted on a small corpus of 780 Ofsted reports (n=2, 639, 523 tokens), which aimed to compare constructions of school leadership across two Ofsted frameworks: the 2019 Education Inspection Framework (EIF) and its predecessor, the 2015 Common Inspection Framework (CIF). The corpus size was limited by the number of schools awarded the highest ('Outstanding') grade during the years 2015-2019. The stated aim of the EIF was to provide a more holistic insight into schools' offer to their community. However, responses to the EIF were generally negative.

Supporting and extending research which has explored headteachers' negative responses to the EIF, the corpus analysis presented indicates that the 2019 EIF constructs increased ambiguity and complexity around the nature of a good curriculum, which has the effect of placing additional responsibilities upon school leaders. Furthermore, corpus analysis suggests that the 2019 EIF extends school leaders' responsibilities regarding the mental health of teachers working in schools. Given the high-stakes nature of school inspection in England, I argue that the discursive patterns identified in this research are likely to have engendered material and affective consequences for school leaders, and are therefore of great importance to education policymakers.

List of references

Ofsted (2015) Common Inspection Framework https://www.gov.uk/government/publications/common-inspection-framework-education-skills-and-early-years-from-september-2015

Ofsted (2019) Education Inspection Framework https://www.gov.uk/government/publications/education-inspection-framework

Towards a reliable annotation framework for crisis MT evaluation: Addressing error taxonomies and annotator agreement

Maria Carmen Staiano[1], Lifeng Han[2], Johanna Monti[3], Chiusaroli Francesca[1]

[1]University of Macerata; [2]Leiden University; [3]University of Naples L'Orientale

Accurate human annotations are essential for evaluating the quality of machine translation (MT) outputs, particularly in sensitive contexts such as crisis communication.

This paper presents a detailed analysis of the annotation process used in the ITALERT (Italian Emergency Response Text) corpus, specifically designed to evaluate the performance of neural machine translation (NMT) systems and large language models (LLMs) in translating high-stakes messages from Italian to English.

The study is guided by the following research questions (RQs):

RQ1: Do existing MT error taxonomies adequately reflect the key features of crisis communication?

RQ2: Can we design an improved annotation framework that integrates decision-support tools and promotes consistency and reliability in crisis-related MT evaluation?

The methodology involved corpus compilation, automatic translations using Google Translate and ChatGPT-4, and human quality assessment through manual annotation.

An initial draft of annotation guidelines was produced to ensure a shared understanding and consistent application of error categories. Following a preliminary annotation phase, ambiguous cases were collected and resolved through structured discussions. To validate annotation reliability, we measured inter-annotator agreement (IAA) using established metrics in computational linguistics (Artstein and Poesio, 2008) such as Cohen's Kappa (Cohen, 1960), Fleiss' Kappa (Fleiss, 1971) and Krippendorff's Alpha (Krippendorff, 2011) for multi-annotator agreement.

Results showed strong agreement overall, though slightly higher consistency was noted for Google Translate (Fleiss' κ = 0.82, Krippendorff's α = 0.83) compared to ChatGPT (Fleiss' κ = 0.78, Krippendorff's α = 0.79). Pairwise analysis highlighted variations in agreement across annotators and MT outputs, revealing specific challenges in annotating different MT outputs.

These findings emphasize the importance of rigorous annotation procedures and demonstrate that clarity in guidelines and structured decision-support tools significantly improve inter-annotator reliability. The outcomes offer valuable methodological insights for corpus annotation within crisis contexts, underscoring the need for domain-sensitive training and robust, well-defined annotation protocols.

List of references

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. Computational linguistics, 34(4), 555-596.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and psychological measurement, 20(1), 37-46.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. Psychological bulletin, 76(5), 378.

Krippendorff, K. (2011). Computing Krippendorff's alpha-reliability.

Examining intraregister variation using cluster analysis

Shelley Staples

University of Arizona

Corpus research often compares texts across groups of predefined, socially constructed categories, e.g., registers or communicative purposes. However, we also know there can be extensive linguistic variation within categories that is important to explore (Egbert et al., 2024). Cluster analysis facilitates the identification texts that show similar use of linguistic variables, regardless of their predefined categorization, allowing for quantitative examination of variation within registers or communicative purposes (Staples & Biber, 2015).

The current study examines texts from a corpus of student writing to explore variation within the categories of communicative purpose and assignment. The corpus consists of 8,343 texts, classified into 34 assignment types (based on rubrics) and eight communicative purposes (based on Nesi & Gardner's 2012 Genre Families). The corpus is first analyzed using multidimensional analysis (MDA), a method to identify constellations of linguistic features that perform particular functions. The MDA revealed five dimensions, which then served as the basis for clustering cases within the corpus.

The results reveal four distinct clusters: 1) Strong Impersonal, Informational Discourse; 2) Strong Personal Stance-driven Discourse and Analysis of (Past) Actions; 3) Mixture of Personal Stance-driven Discourse, Analysis of (Past) Actions, and Forecasting of Possibilities; 4) Strong Impersonal, Informational Discourse and Informationally Dense Summaries. Further, the results show that certain assignment types/communicative purposes (e.g., Literacy Narratives, classified as Narrative Recounts) show little variation (representing Cluster 2). However, other assignment types/communicative purposes (e.g., Annotated Bibliographies, classified as Essays) were spread evenly across Cluster 1 and 4, suggesting that while all of these texts contained Impersonal, Informational Discourse, some focused more on summarizing and thus were more aligned with Literature Surveys (e.g., Literature Review assignments). These results have important implications for the teaching and evaluation of these assignments and identify places where students can make informed choices for language use based on larger communicative goals.

List of references

Egbert, J., Biber, D., Keller, D., Gracheva, M. (2024). Register and the dual nature of functional correspondence: accounting for text-linguistic variation between registers, within registers, and without registers. Corpus Linguistics and Linguistic Theory, 20(3): 505-538.

Staples, S. & Biber, D. (2015). Cluster analysis. In L. Plonsky (Ed.), Advancing quantitative methods in second language research (pp. 243-274). Routledge.

Nesi, H., & Gardner, S. (2012). Genre across the disciplines: Student writing in higher education. Cambridge University Press.

Corpora and statutory language: How techniques from corpus linguistics can assist statutory interpretation and contribute to improving clarity in legislation

Samantha Stark

Lancaster University

The rule of law is an important constitutional principle in the UK. It requires that laws are accessible, intelligible, clear and predictable (Prentis, 2023) so that they work to guide people's behaviour, which is their primary function. The words used in legislation therefore need to be clear and in accordance with how the public use and understand those words. This is of particular importance in relation to serious criminal offences such as terrorism, for which custodial sentences are frequently imposed upon conviction.

The UK definition of terrorism is very broad (Hall, 2022). Some words contained in the definition have been analysed in the courts, so their meaning is clearer. Others have not, such as *influence* in the phrase *influence the government*. As the definition does not specify that the influence needs to be negative or untoward, the question then is, does *influence* have a negative prosody in everyday or legal language, so that the negativity can be inferred? Corpus linguistics techniques, such as concordance analyses and Word Sketches, can help to answer this question.

In this paper, I will demonstrate how corpora can be used to analyse legislation and identify where slight amendments could be made to make the law clearer in accordance with the rule of law. Using the English Web 2021 corpus and a custom-built criminal law corpus, I will focus on words in the terrorism definition that could impact the breadth of the definition. Regarding *influence*, my analysis has shown it has a neutral prosody in general English, meaning any kind of influence, even that considered normal in a democratic society, could be sufficient to satisfy the legal definition of terrorism.

Delegates will come away with new insights into the relationship between law and linguistics, and how linguistic perspectives can contribute to upholding the rule of law.

List of references

Prentis KC (2023), Keynote speech on the Rule of Law, accessible from https://www.gov.uk/government/news/attorney-general-delivers-speech-on-the-rule-of-law#:~:text=The%20rule%20of%20law%20is,to%20courts%20that%20are%20independent. [accessed 08.07.2024]

Hall KC (2022) The Terrorism Acts in 2020, accessible from https://terrorismlegislationreviewer.independent.gov.uk/wp-content/uploads/2022/04/Terrorism-Acts-in-2020.pdf [accessed 17.01.2025]

Reading in a second language: How do adapted Czech texts differ from originals?

Zaneta Stiborska, **Michaela Nogolova**

University of Ostrava

This corpus study explores the syntactic complexity of adapted literary texts for learners of Czech as a second language (L2). It aims to evaluate how syntactic complexity of texts varies across language proficiency levels (A2, B1, B2) as defined by the Common European Framework of Reference for Languages (CEFR 2001), and how it varies between adapted and original texts.

The Czech corpus consists of ten adapted books specifically designed for L2 learners. A quantitative analysis of syntactic complexity was conducted using metrics such as Average Sentence Length (in words and clauses), Average Clause Length (in words), Mean Dependency Distance and Mean Hierarchical Distance (Lu 2010 & 2011; Liu 2008). The results reveal a progressive increase in syntactic complexity from A2 to B2 levels, indicating a correlation between the linguistic features of adapted texts and their intended proficiency level. Furthermore, original texts are syntactically more complex across nearly all analysed metrics. The difference between adapted and original texts narrows as the proficiency level increases.

These findings contribute to a deeper understanding of text design for second language learners and highlight the importance of syntactic calibration in educational materials. Future research may further explore the implications for teachers and textbook authors, particularly in optimizing reading materials for language acquisition at various proficiency levels.

List of references

Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. Journal of Cognitive Science, 9(2), pp. 159–191.

CEFR: Council of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division. (2001). Common European framework of reference for languages: Learning, teaching, assessment. Cambridge University Press.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. International journal of corpus linguistics, 15(4), 474–496.

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. TESOL quarterly, 45(1), 36–62.

The unique style of false narratives: Mapping the anti-system media register using quantitative methods

Konstantin Sulimenko

Faculty of Arts, Charles University

Employing multidimensional analysis (MDA; Biber, 1988), we describe the register of Czech anti-system media—i.e., media spreading (Russian) propaganda, false or rigged narratives, and/or fake news—for a comprehensive understanding of anti-system disinformation techniques and strategies. We propose that our findings effectively prove the potential for the existence of a unique anti-system media register in many other languages.

Although new methods of analyzing the language of fake news are regularly presented, they usually focus only on two key issues: content analysis, including keyword analysis (Cvrček and Fidler, 2022), topic modelling (Boberg et al., 2020), and qualitative approaches (Trnavac and Põldvere, 2024), as well as automatic fake news detection through linguistic features (see Lugea 2021 for an overview). However, we argue that in the study of disinformation, linguistic form is equally significant to content (cf. Grieve and Woodfield, 2023). To address this gap, we provide a detailed analysis and evaluation of the linguistic strategies used to present fake narratives to readers.

Using the additive MDA approach (Berber Sardinha et al., 2019; building on a model by Cvrček et al., 2021), we analyzed a corpus of 5,905 articles from January 2022 published in Czech mainstream and anti-system media, mapping linguistic features onto an 8-dimensional space.

The results proved that there are significant differences between the mainstream and anti-system journalistic registers, as the anti-system register is more spontaneous (dimension 2), cohesive (dimension 3), monothematic (dimension 4), general (dimension 6), attitudinal (dimension 8), and contains a higher amount of addressee coding (dimension 5).

Interpreting the results, we identified two key strategies of anti-system authors. First, colloquialism and informality, resembling the style of Facebook posts, appear to be used to attract readers. Second, anti-system media present fewer concrete pieces of information, focusing more on their own contextualization and interpretation of the already known facts and events.

List of references

Berber Sardinha, T., Pinto, M. V., Mayer, C., Zuppardi, M. C., & Kauffmann, C. H. (2019). Adding registers to a previous multi-dimensional analysis. In Biber, D., & Egbert, J. (Eds.), Multi-dimensional analysis: Research methods and current issues, 165–186.

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

Boberg, S., Quandt, T., Schatto-Eckrodt, T., & Frischlich, L. (2020). Pandemic populism: Facebook pages of alternative news media and the corona crisis--A computational content analysis. arXiv preprint arXiv:2004.02566.

Cvrček, V., Komrsková, Z., Lukeš, D., Poukarová, P., Řehořková, A., & Zasina, A. J. (2021). From extra-to intratextual characteristics: Charting the space of variation in Czech through MDA. Corpus Linguistics and Linguistic Theory, 17(2), 351–382.

Cvrček, V., & Fidler, M. (2022). No Keyword is an Island: In search of covert associations. Corpora, 17(2), 259-290.

Grieve, J., & Woodfield, H. (2023). The Language of Fake News. Elements in Forensic Linguistics, Cambridge University Press.

Lugea, J. (2021). Linguistic approaches to fake news detection. In Deepak, P., Chakraborty, T., Long, C., & Santhosh, K. G. (2021). Data Science for Fake News. Springer: Cham.

Trnavac, R., & Põldvere, N. (2024). Investigating Appraisal and the Language of Evaluation in Fake News Corpora. Corpus Pragmatics, 8(2), 107-130.

Agency, vulnerability, and relational identity in online suicide notes: A personal pronoun and n-gram analysis

Kexin Tan, Bernard Ayine

The Hong Kong Polytechnic University

Traditionally confined to private physical documents, suicide notes have increasingly emerged in digital spaces as platforms where suicidal individuals publicly articulate distress, seek connection, and negotiate self-representation (Fernández-Cabana et al., 2015; Durkee et al., 2011). This shift in medium necessitates a deeper understanding of how suicidal individuals construct their identities in these digital environments. The study extends prior research by examining how suicidal individuals linguistically construct identity and relational dynamics in a 530,000-word corpus of digital suicide notes (suicideproject.org, 2016–2024) by employing corpus-based methods such as N-gram analysis and semantic network mapping to analyze personal pronoun usage (I, me, we, us, you, they, he, she, him, her, them). Findings reveal that agentive pronouns (I, we) oscillate between assertiveness and vulnerability, while relational pronouns (you, they) frequently denote blame or disconnection, clustering around adversarial dynamics. Contextual contrasts emerge: farewell messages correlate with positive we-collocates like “love” and “share,” whereas justifications feature negative I-collocates such as “disgust” and “failure.” Network mapping further illustrates fragmented communal bonds (we, us) and adversarial relationships (they, you). These patterns advance understanding of digital self-representation in crises, highlighting how pronoun use mediates emotional expression and social ties. The study informs targeted mental health interventions by identifying linguistic markers of isolation or despair in online spaces. By bridging linguistic insights with mental health research, this work contributes to nuanced strategies for supporting at-risk individuals in digital communities. This study strictly adheres to ethical guidelines, prioritizing anonymization and sensitivity in analyzing suicide discourse.

List of references

Durkee, T., Hadlaczky, G., Westerlund, M., & Carli, V. (2011). Internet pathways in suicidality: a review of the evidence. International Journal of Environmental Research and Public Health, 8(10), 3938-3952.

Fernández-Cabana, M., Ceballos-Espinoza, F., Mateos, R., Alves-Pérez, M. T., & García-Caballero, A. A. (2015). Suicide notes: clinical and linguistic analysis from the perspective of the Interpersonal Theory of Suicide. The European Journal of Psychiatry, 29(4), 293-308.

Robinson, J., Cox, G., Bailey, E., Hetrick, S., Rodrigues, M., Fisher, S., & Herrman, H. (2016). Social media and suicide prevention: a systematic review. Early Intervention in Psychiatry, 10(2), 103-121.

Zaśko-Zielińska, M. (2022). The linguistic analysis of suicide notes. In Language As Evidence: Doing Forensic Linguistics (pp. 373-417). Cham: Springer International Publishing.

Parenting or partnership? Corpus-assisted analysis of institution-people dynamics in university sustainability reporting

Yingnian Tao

Lancaster University

This study analyses the sustainability reports published by 46 UK universities, focusing on how institutions construct their image in relation to their staff, students, and community members. Prior research on sustainability discourse in higher education has largely centred on institutional image construction, with limited attention to how universities position themselves in relation to their stakeholders. This study addresses this gap by exploring the tensions between institutions and their staff, students and community members, highlighting the dynamics of agency, participation, and image-making in sustainability communication in higher education.

Using corpus linguistic techniques, we compiled a dataset of 1,398,916 tokens across 158 documents from UK universities identified in the 2024 QS Sustainability Rankings. Collocation and concordance analyses were performed using Sketch Engine to compare representations of "we/university" (institution) versus "staff/students/community" (people).

Findings reveal an imbalanced image construction. Universities are consistently presented as active agents driving sustainability changes in campus operations (e.g., biodiversity, recycling), curriculum, research activities, and community engagement. These actions construct a caring, prestigious institutional persona, characterised by keen awareness and proactive and top-down approach to climate action. In contrast, staff, students, and community members are often depicted as passive recipients of the university's facilities and services who have little agency over these initiatives and programmes. Sometimes, they are used to justify shortcomings in university climate performances.

This "parenting" narrative positions universities as providers and decision-makers, while downplaying the critical contributions of their people. This framing diminishes collective agency between institution and its stakeholders and may discourage stakeholders from undertaking further environmental actions.

This study contributes to critical discourse analysis of sustainability communication in higher education and advocates for participatory narratives. It challenges institutions to move beyond hierarchical narratives in sustainability reporting towards acknowledging the existing efforts of its people and empowering these stakeholders to achieve collaborative and inclusive sustainability practices.

Acknowledgement
This study is supported by the Wellcome Trust (228123/Z/23/Z).

Introducing NomadLingo: A corpus of translingual spoken interactions in European digital nomad communities

Novella Tedesco, Silvia Bernardini, Cristiana Cervini

University of Bologna

This paper introduces NomadLingo, a corpus containing transcripts of naturally occurring spoken interactions recorded during informal social events in European digital nomad communities. Developed within the Ph.D. project Fluid Languages Observatory (FLO, thenomadlinguist.eu/flo), the project applies innovative methods to corpus design, construction, and annotation. Digital nomad communities, groups of remote workers adopting a traveling lifestyle and gathering in coworkings and colivings (Woldoff & Litchfield, 2021), exemplify culturally diverse, highly digitalized, mobile settings where communication assumes fluid forms and translanguaging mechanisms become particularly evident (Garcìa and Wei 2014, Canagarajah, 2017).

The corpus comprises 12 hours of transcribed audio from 50 participants, recorded in Madeira and Canary Islands in 2024. It documents communication in English as a Lingua Franca (Mauranen and Ranta 2009), enriched by trans/plurilingual practices like intercomprehension (Capucho 2017). Data collection followed an ethnographic approach (Tusting 2020) and received approval by the Bioethical Committee at the University of Bologna in 2023.

The paper details methodological innovations in data collection, processing, and analysis, including semi-automatic transcription using Open AI Whisper and transcription conventions adapted from Jefferson (1984) to represent salient features of spoken communication, such as tone changes, pauses, and overlaps. Contextual metadata in .csv format enriches the corpus, while details about speakers' nationality and languages are annotated in-text.

Inspired by translanguaging theories (Garcìa and Wei 2014), a simple annotation system highlights segments where translanguaging becomes evident, i.e. instances of code-switching, code-mixing, translation. The XML-based encoding of both contextual and analytical annotation follows FAIR principles to ensure reusability and accessibility[1]. This allowed for a first analysis with results suggesting positive effects of translingual practices on fluency.

[1] At the time of writing, a demo version is downloadable from thenomadlinguist.eu/nomadlingo/. However, NomadLingoSmall is planned to be deposited and available for consultation on NoSketchEngine before the end of April 2025.

List of references

Canagarajah, A. Suresh, ed. 2017. The Routledge Handbook of Migration and Language. New York: Routledge.

Capucho, Filomena. 2017. "Interactions Professionnelles Plurilingues: Entre Intercompréhension et Interproduction." Repères DoRiF 12. Rome: Do.Ri.F.-Università.

García, Ofelia, and Wei Li. 2014. Translanguaging: Language, Bilingualism and Education. London: Palgrave Macmillan.

Jefferson, G. (1984). Transcript notation. In J. M. Atkinson & J. Heritage (Eds.), Structures of Social Action: Studies in Conversation Analysis (pp. ix–xvi). Cambridge: Cambridge University Press.

Mauranen, Anna, and Elina Ranta, eds. 2009. English as a Lingua Franca: Studies and Findings. Newcastle upon Tyne: Cambridge Scholars Publishing.

Tusting, K. (2020). The Routledge Handbook of Linguistic Ethnography. Routledge.

Woldoff, R. A., & Litchfield, R. C. (2021). Digital Nomads: In Search of Meaningful Work in the New Economy. New York: Oxford University Press.

Discursive practices of top travel influencers: Results from a small-scale corpus of Instagram posts

Emanuela Tenca

University of Milan

Contents published by travel influencers online are a popular source of information for people planning their holidays. Travel influencers use their profiles as a source of income (Duffy and Kang 2020; Willment 2020), and to present their story as authentic, they harness travel discourse strategies; however, they also apply the promotional strategies of tourism discourse to raise the visibility of their brand (Azariah 2017).

Following from these premises, this study analyses how top travel influencers emotionally connect with their audiences and enhance their social media presence by answering these questions:

1. What type of linguistic strategies are applied for engaging followers?
2. How is the uniqueness of the influencers' brands discursively expressed?

To answer these questions, a small-scale study was conducted on a self-compiled corpus including 300 Instagram posts published by five top travel influencers. The corpus (38,763 tokens; 32,504 words) was uploaded to Sketch Engine (Kilgariff et al. 2014), and divided into five sub-corpora, one for each influencer. Wordlists were run for each sub-corpus to identify frequent features of travel or of tourism discourse. Next, keywords were extracted initially by setting the small-scale corpus as a focus corpus and EnTenTen21 as a reference corpus; subsequently, each sub-corpus was compared against the other four. Concordance lines were finally examined to understand how frequent words and keywords are used by each influencer.

The findings show how ego-targeting (Dann 1996) is implemented to include followers in the travel experience, while targeting them as potential customers of travel services. The keywords revealed tourist destinations and transportation as two key topics. Differences emerged in how influencers deal with these topics, namely by adopting emotively connoted language, or by integrating technical terminology. These differences depend on the brand personality staged by each influencer, which may rely on ideas of authenticity and self-fulfilment, or of competence and professionalism.

List of references

Azariah, D. R. (2017). Tourism, travel, and blogging. A Discursive analysis of online travel narratives. London: Routledge.

Dann, G. M. S. (1996). The Language of Tourism: A Sociolinguistic Perspective. CAB International: Wallingford.

Duffy, A. and Kang H. Y. P. Kang (2020). "Follow me, I'm famous: Travel bloggers' self-mediated performances of everyday exoticism", Media, Culture, & Society 42 (2), 172-190.

Kilgarriff, A et al. (2014). "The Sketch Engine: Ten years on", Lexicography, Journal of Asialex 1 (1), 7-36.

Willment, N. (2020). The travel blogger as digital nomad: (Re-)imagining workplace performances of digital nomadism within travel blogging work. Information Technology & Tourism, 22, 391-416.

The discourse of wellbeing in American late modern self-help texts for teachers and students (1830-1930): Findings from a small-scale study

Emanuela Tenca, Andrea Nava, Luciana Pedrazzini

University of Milan

Wellbeing has been investigated in various research fields, including education (Kern and Wehmeyer 2021). The historical development of this concept can be analysed in the discourse of texts published in times of major changes, e.g. the Late Modern period. This is when a new genre of popular non-fiction, self-help books, emerged in the US as a tool for self-improvement (Alharbi 2023).

Our aim is to investigate how the concepts of self-help and wellbeing are interrelated in texts for American teachers and students published between the 1830s and the 1930s. Our investigation is informed by these questions:

1) How is wellbeing discursively constructed in self-help texts for the general public versus texts targeting teachers and students?

2) Which key themes underlie the discourse of wellbeing in these texts?

3) Which rhetorical devices are adopted to motivate readers to do and feel well?

To answer these questions, we conducted a small-scale study on a self-compiled corpus (712,609 tokens; 610,064 words) consisting of three sub-corpora (general, students, and teachers). The corpus was uploaded to Sketch Engine (Kilgarriff et al. 2014) and analysed through a mixed methods approach. Wordlists were run to retrieve the most frequent words in each sub-corpus. Keywords were extracted to identify items unique to the discourse of teachers' and students' self-help texts by comparing the two sub-corpora against the general sub-corpus. Word sketches helped explore the collocational behaviour of the keywords. Finally, the patterns around those keywords were analysed qualitatively using the concordance tool.

The results show that wellbeing is related to concepts of self-reliance, self-improvement, and achievement. The register varies from authoritative to instruct readers, to more conversational to support self-reflection. While the discourse of wellbeing in texts for students appears to be similar to that of general texts, in texts for teachers it foregrounds challenges in classroom management and teacher-student relationship.

List of references

Alharbi, A. N. (2023). "The rise, definition, and classification of self-help literature". Americana: The Journal of American Popular Culture (1900 to present) 22 (1).

Kern, M. L. and Wehmeyer, M. L. (2021). The Palgrave Handbook of Positive Education. Palgrave Macmillan: Cham.

Kilgarriff, A et al. (2014). "The Sketch Engine: Ten years on". Lexicography, Journal of Asialex 1 (1), 7-36.

Linguistic mechanisms of trust in the Pick-Up artist community: A multidimensional analysis

Annina Van Riper

University of Birmingham

The Pick-up Artist (PUA) community is an online community of (mostly) men who are dedicated to the art of seducing women. Community leaders, so-called “seduction experts”, engage with novice PUAs through two primary genres: lectures and how-to videos. Lectures are offline, in-person trainings and how-to videos are influencer style videos posted on social media; the communicative purpose of both genres is to train PUAs on how to sway women to be agreeable to a PUAs desires (Dayter & Rüdiger, 2019). This paper uses a corpus of transcribed lectures and how-to videos to investigate the discursive mechanisms of trust. Building trust is critical to PUA community leaders as it underpins their credibility and authority among followers. To study how trust is manifest, this research employs multidimensional analysis (MDA) (Biber, 1988) to identify and compare discursive features of each aforementioned genre.

The analysis presented addresses the following key research questions: (1) what linguistic features are prominent in establishing trust? (2) How do these features differ between lectures and how-to videos? (3) How do these patterns contribute to understanding the rhetorical strategies of trust building in the PUA community? Preliminary findings suggest that lecture videos can be characterised by informal narration whereas how-to videos comprise of interactive features such as direct address and imperatives. I argue that MDA is a suitable method despite its novel use investigate trust as it can identify genre-specific differences which highlight the interplay of linguistic features and their relevance to trust-building resources. Moreover, MDA allows for the identification of significant PUA genre features thus providing comprehensive insight on a community known for promulgating misogynistic ideologies and violence towards women.

List of references

Biber, D. (1988). Variation across Speech and Writing. Cambridge: Cambridge University Press.
Dayter, D., & Rüdiger, S. (2022). The Language of Pick-Up Artists: Online Discourses of the Seduction Industry (1st ed.). Routledge. https://doi.org/10.4324/9781003041313

Variation in topic development and register across modalities: A corpus-based study of speaking and writing tasks

Odette Vassallo, **Geraldine Mark**

University of Malta

In this paper we first describe the design and creation of a corpus of 14-25 year-old Maltese users of English. We explore topic mapping and register difference in the developing linguistic repertoires. Maltese and English are the two official languages of Malta and exposure to both languages differs in great measure. The corpus reflects stages of education, diverse educational scenarios, age range, and linguistic attainment, across individuals and groups who are multilingual to varying degrees.

We focus on the creation and analysis of two subcorpora which represent the speaking and writing of (1) an entire population of 14-15 years olds and (2) university level students across disciplines. We showcase how data was gathered using an integrated approach with tasks designed to encourage students to engage in spoken interaction within a group and writing based on ideas derived from the speaking. Thus, creating a bridge between topics generated during the speaking task and writing produced immediately after the group interaction. These tasks move away from the traditional curricular framework as they incorporate an interactional element to potentially scaffold the writing process and product. Each task was situated within a specific genre to motivate students to generate distinct characteristics.

In investigating the data, we take a mixed-methods approach, exploring how language usage varies across age groups, task, interaction, genre and register. We show how sensitivity to the needs of the task plays out across the data in the overall development of register awareness. A case study drawn from the corpus illustrates how specific tasks highlight the contrasts between how students orient their speaking and writing towards topic development and text constructions. Findings reveal variations in topic mapping, how one modality informs and supports the other, as well as the role of interaction and register in shaping linguistic choices across modalities.

Language ideologies and language practices: A corpus-assisted discourse studies approach to the (re)negotiation of multilingualism policy at the United Nations

Rachelle Vessey[1], **Lisa McEntee-Atalianis**[2]

[1]Carleton University; [2]Birkbeck, University of London

Multilingualism plays a central albeit contested role in the United Nations, with conflicting organizational and national goals and priorities being advanced by members (McEntee-Atalianis and Vessey, 2020, 2024). In 1995, a multilingualism resolution was adopted by vote, but has continued to be debated and redrafted. Using these debates and voting records as data, we compare and contrast linguistic choices (choice of official language), metadiscursive constructions (representations of multilingualism), and metapragmatic actions (voting records) using corpus-assisted discourse studies (CADS), explicating the role of language and its national context following an accountability framework (Bednarek, Schweinberger and Lee, 2024).

The corpus consists of 57,499 words in English and translations of the six official languages of the United Nations. 23 unique subcorpora were created to account for voting records and official languages used. Adapting an approach used by Potts and Weare (2018), we tagged first person plural pronouns for “footing” (Goffman, 1981) and “positioning” (Davies and Harré, 1990). Tags accounted for (1) the official language used, (2) the position adopted (e.g., advancing national vs. supranational priorities), and (3) the vote on the 1995 resolution. Using SketchEngine, we determined which official languages were used most in subcorpora organized by voting records. Next, tag searches revealed which positions were predominantly adopted by speakers in each case. Finally, all concordance lines of “we” were labelled for verb patterns using systemic functional linguistics.

Findings revealed that the “Yes” vote was primarily articulated in French, where speakers adopt a United Nations positioning and modal verbs to call for action. The “No” vote was articulated exclusively in English and “No” voters position themselves as national representatives, advancing national interests, using mental processes to express concerns about the Resolution. Our approach highlights the holistic insights about language ideologies and language practices that can be gleaned using a corpus-assisted discourse studies approach.

List of references

Bednarek, M., Schweinberger, M., and Lee, K. K. H. (2024). Corpus-based discourse analysis: from meta-reflection to accountability. Corpus Linguistics and Linguistic Theory, 20 (3) 1-28. https://doi.org/10.1515/cllt-2023-0104

Davies, B., & Harré, R. (1990). Positioning: The discursive production of selves. Journal for the Theory of Social Behaviour, 20(1), 43-63.

Goffman, E. (1981). Forms of Talk. University of Pennsylvania Press.

McEntee-Atalianis, L. and Vessey, R. (2020). Mapping the language ideologies of organisational members: a Corpus Linguistic Investigation of the United Nations’ General Debates (1970-2016). Language Policy, 19, 549-573.

McEntee-Atalianis, L. and Vessey, R. (2024). Using corpus linguistics to investigate agency and benign neglect in organisational language policy and planning: the United Nations as a case study. Journal of Multilingual & Multicultural Development, 45 (2), 358-373.

Potts, A. and Weare, S. (2018). Mother, Monster, Mrs, I: A critical evaluation of gendered naming strategies in English sentencing remarks of women who kill. International Journal for the Semiotics of Law, 31(1), pp. 21-52. (10.1007/s11196-017-9523-z)

Using large comparable corpora for pharmaceutical translation: The case of antidepressant medication

Fang Wang[1], **Sabine Braun**[2]

[1]University of Surrey; [2]University of Surrey

Comparable bilingual corpora are collections of texts in two different languages with similar purposes and characteristics, such as text type, content, and domain. Utilizing comparable corpora in translation training offers numerous benefits, including the identification of terminological and referential forms of equivalence, the construction of knowledge, and the production of Target Texts (TT). In recent years, there has been a growing demand for English-Chinese pharmaceutical translation due to China's emergence as one of the world's largest markets for prescription drugs. This project has developed a substantial comparable corpus (9,757,791 words) focused on antidepressants to enhance students' training in pharmaceutical translation. The corpus consists of two components: first, antidepressant advertisements for Fluoxetine, Duloxetine, Venlafaxine, Citalopram, and Sertraline that are recorded in official British and Chinese prescription drug databases; and second, articles from medical journals that feature the aforementioned five antidepressants, sourced from Medline and CNKI (China National Knowledge Infrastructure). In training students, the corpora are utilized in several key ways. Firstly, they are employed for terminology extraction and comparison, particularly focusing on terminological collocability. Secondly, they aid in knowledge construction regarding the indications, side effects, and mechanisms of action of antidepressants, along with drug interactions and dosage administration in clinical contexts. Thirdly, they help students develop an understanding of stylistic nuances in pharmaceutical advertisements and academic medical journal articles. Examples will be provided to support the above points. By honing their skills in these areas, students are able to attain a comprehensive and accurate understanding of pharmaceutical texts on antidepressants in English and Chinese, enabling them to produce high-quality translations in this specialized field. Trainers play a crucial role in guiding students through the process of creating glossaries and producing TTs, gaining insights into their learning and cognitive processes, and providing advanced levels of supervision and support.

Acknowledgement
This research is funded by British Academy/Leverhulme Small Research Grants (reference: SG2122\210988).

The role of corpus-based discourse analysis in supporting interpreter preparation for maritime conferences

Fang Wang[1], Huiling Liu[2], Gang Zeng[3]

[1]University of Surrey; [2]Guangzhou Maritime University; [3]Dalian Maritime University

To interpret for highly specialised maritime conferences, interpreters need to acquire in-depth understanding of subject knowledge surrounding key topics in maritime discourse. This paper analyzes the key topics addressed in conferences organized by the International Maritime Organization (IMO) from 2018 to 2024, with the objective of enhancing the subject knowledge of interpreter trainees within the maritime community, preparing them for their booth practice at the IMO on an annual basis. To achieve this aim, we compiled a comprehensive corpus encompassing all summaries of meetings conducted by the IMO Assembly, the IMO Council, and twelve IMO (sub)committees, including the Marine Environment Protection Committee (MEPC) and the Maritime Safety Committee (MSC), among others. The resulting corpus contains approximately 10 million words.

Utilizing the Sketch Engine, we identified the most frequently occurring words and phrases that represent significant topics within the corpus. Collocation and concordance analyses were conducted around these key terms, which include "shipping," "maritime," "safety," "member states," "sustainable development," and "international shipping." Furthermore, critical discourse analysis was employed to interpret linguistic phenomena through the framework of actual maritime contexts.

The findings of this study assist students in understanding the principal topics discussed within the maritime community, thereby enabling them to grasp overarching subject knowledge and familiarize themselves with the typical usage of maritime English essential for their specialized interpreting tasks. This research exemplifies how trainers in translation and interpreting can effectively bridge the gap between corpus linguistics and classroom instruction in English for Specific Purposes (ESP). Our specialized corpus in marine English has proven invaluable in providing students with rich contextual information pertinent to maritime conferences. Investigating the distribution of the most frequent collocations associated with key concepts in the maritime domain aids students in developing a deeper understanding of the language applicable to these specific contexts.

A diachronic analysis of stance markers in Chinese sustainability reports: A corpus- and LLM-based approach

Mingzhu Wang, Chengyu Alex Fang

City University of Hong Kong

This study investigates stance markers in sustainability reports within China's real estate sector. Stance markers, or evaluative markers, signal the speaker's or writer's attitude toward a topic (Biber & Finegan, 1988; Hunston & Thompson, 2000). Despite the critical role of stance markers in shaping corporate image (Fuoli, 2018), the diachronic changes in their usage within specific industries—particularly how they vary between periods of boom and recession—remain underexplored. By employing corpus-based methods, we built a corpus of 1,193,743 word tokens from 70 English reports produced over the period of 2016–2023, sourced from 10 real estate companies in China. Using the most advanced large language model GPT-4o, we refined previous classifications of stance markers by incorporating industry-specific terms (e.g., *net-zero*, *supply chain resilience*), prepositional phrases (e.g., *without a doubt*), and passive constructions (e.g., *been granted*). This approach resulted in an F1 score of 0.8. Notably, for modal and semi-modal verbs and stance adverbs, they are 0.93 and 0.90, respectively. Our findings reveal that most companies increased attitudinal stance markers from 2016 to 2020, followed by a sharp decrease from 2021 to 2023. Attitudinal verbs, adverbs, and evaluative adjectives were the most frequently used categories. Evaluated items such as *committed*, *aims*, *responsive*, and *strictly* were prevalent, reflecting a strategic emphasis on future intentions and obligations to meet policy requirements and societal expectations. A corpus word cloud analysis highlights themes such as *supply chain*, *employee*, and *environment management*, aligning with corporate communication strategies aimed at enhancing corporate image. While this research is limited by its focus on English reports in a highly specialised industry sector, it offers valuable insights into linguistic shifts in business communication as an under-studied register. The findings provide broader implications for understanding corporate discourse and increasing the transparency of sustainability reports.

List of references

AIEZZA, M. C. 2015. "We may face the risks"…"risks that could adversely affect our face." A corpus-assisted discourse analysis of modality markers in CSR reports. Studies in Communication Sciences, 15, 68-76.

AL-SHUNNAG, M. 2014. Stance in political discourse: Arabic translations of American newspaper opinion articles on the Arabs Spring', University of Salford (United Kingdom).

ARRESE, J. I. M. N. 2009. Effective vs. epistemic stance, and subjectivity/intersubjectivity in political discourse. A case study. Studies on English modality in honour of Frank Palmer. Linguistic Insights, 111, 23-131.

BIBER, D. 1993. Representativeness in corpus design. Literary and linguistic computing, 8, 243-257.

BIBER, D. 2006a. Stance in spoken and written university registers. Journal of English for Academic Purposes, 5, 97-116.

BIBER, D. 2006b. University language: A corpus-based study of spoken and written registers. John Benjamins.

BIBER, D. & FINEGAN, E. 1988. Adverbial stance types in English. Discourse processes, 11, 1-34.

BIBER, D. & FINEGAN, E. 1989. Styles of stance in English: Lexical and grammatical marking of evidentiality and affect. Text-interdisciplinary journal for the study of discourse, 9, 93-124.

BIBER, D., JOHANSSON, S., LEECH, G., CONRAD, S. & FINEGAN, E. 2000. Longman grammar of spoken and written English. Longman London.

CROSTHWAITE, P., BOYNTON, S. & COLE III, S. 2017. Exploring rater conceptions of academic stance and engagement during group tutorial discussion assessment. Journal of English for Academic Purposes, 28, 1-13.

FILIENKO, D., WANG, Y., JAZMI, C. E., XIE, S., COHEN, T., DE COCK, M. & YUWEN, W. 2024. Toward Large Language Models as a Therapeutic Tool: Comparing Prompting Techniques to Improve GPT-Delivered Problem-Solving Therapy. arXiv preprint arXiv:2409.00112.

FUOLI, M. 2018. Building a trustworthy corporate identity: A corpus-based analysis of stance in annual and corporate social responsibility reports. Applied linguistics, 39, 846-885.

GRAY, B. & BIBER, D. 2012. Current conceptions of stance. Stance and voice in written academic genres, 15-33.

GROUP, P. R. 2016. Not so 'innocent'after all? Exploring corporate identity construction online. Discourse & Communication, 10, 291-313.

HALLIDAY, M. A. K. & MATTHIESSEN, C. M. 2013. Halliday's introduction to functional grammar, Routledge.
HUNSTON, S. & THOMPSON, G. 2000. Evaluation in text: Authorial stance and the construction of discourse: Authorial stance and the construction of discourse, Oxford University Press, UK.
HYLAND, K. 2005. Stance and engagement: A model of interaction in academic discourse. Discourse studies, 7, 173-192.
KHAN, H. U. Z. 2010. The effect of corporate governance elements on corporate social responsibility (CSR) reporting: Empirical evidence from private commercial banks of Bangladesh. International Journal of Law and Management, 52, 82-109.
LIN, Y. 2021. Legitimation strategies in corporate discourse: A comparison of UK and Chinese corporate social responsibility reports. Journal of Pragmatics, 177, 157-169.
LIU, J. & LIU, Q. 2023. Stance constructions in CEO statements of CSR reports of Chinese and US companies. English for Specific Purposes, 70, 237-251.
LUBISA, H., PRATAMA, K., PRATAMA, I. & PRATAMI, A. 2019. A systematic review of corporate social responsibility disclosure. International Journal of Innovation, Creativity and Change, 6.
MIKES, A., OYON, D., JEITZINER, J. & KPMG, G. 2017. Risk management: Towards a behavioral perspective. The Routledge Companion to Behavioral Accounting Research. Oxford: Routledge. doi, 10, 9781315710129-29.
NIELSEN, A. E. & THOMSEN, C. 2007. Reporting CSR–what and how to say it? Corporate Communications: An International Journal, 12, 25-40.
THORNE, S. L. & REINHARDT, J. 2008. “Bridging activities,” new media literacies, and advanced foreign language proficiency. Calico Journal, 25, 558-572.
WU, J. & PAN, F. 2023. Changing patterns of the grammatical stance devices in medical research articles (1970–2020). Journal of English for Academic Purposes, 66, 101305.
ZHANG, Z. 2022. What is ESG Reporting and Why is it Gaining Traction in China? [Online]. China Briefing. Available: https://www.china-briefing.com/news/what-is-esg-reporting-and-why-is-it-gaining-traction-in-china/ [Accessed].

A corpus-based study on the grammaticalization paths of the Chinese classifier 枚 méi

Minyue Wang

Lancaster University

A notable characteristic of Chinese is the utilisation of sortal numeral classifiers, appearing between a number and a noun in modern Mandarin. These classifiers serve the functions of individuating (Aikenvald, 2000) and categorization (Lakoff, 1987). This study aims to investigate the grammaticalization paths of classifiers via quantitative methods. 枚 was chosen for being the most frequent classifier before its replacement 个 gè. The study will mainly focus on the second stage of grammaticalization "from the lexical item to the grammatical item" (Kuryłowicz, 1965).

3,611 instances of 枚 were retrieved from the CCL corpus from the Zhou dynasty to the Republic of China. In addition, instances of 枚 in Modern Chinese were annotated for a synchronic account of gradient usages (Traugott et al., 2007). Manual annotation was conducted on the part of speech and other salient features to prove the changes. The analysis found 26 distinct usages and 19 classifier constructions. Distinctive Collexeme Analysis (Levshina, 2015) was employed to identify constructions that were significantly more or less frequent in certain periods, and to prove the constructional changes, particularly highlighting the competition (Hilpert, 2013) between two similar constructions [Noun Number Classifier] and [Number Classifier Noun]. Utilising this framework, a logistic regression model was constructed to investigate the formal and functional variables in distinguishing the two constructions. Among these, the variables "Dynasty", "Distance" and "Pragmatic" were found to be significant.

The primary accomplishment of this study lies in the utilisation of quantitative methodologies to elucidate the historical transition of two predominant classifier constructions. The study also provides a reasonable explanation of the grammaticalization path of 枚. Informed speculation will be presented for the initial phase and two main parameters for the second stage were identified: the gain of "paradigmaticity" (Lehmann, 1985: 4) and the loss of syntagmatic variability (ibid., p. 5).

List of references

Aikhenvald, A. (2000). Classifiers: A Typology of Noun Categorization Device. Oxford: Oxford University Press.

Hilpert, M. (2013). Constructional change in English: Developments in allomorphy, word formation, and syntax. Cambridge University Press.

Kuryłowicz, J. (1965). The Evolution of Grammatical Categories. Diogenes, 13(51), 55-71.

Lakoff, G. (1987), Women, Fire, and Dangerous Things: What Categories Reveal about the Mind. Chicago: University of Chicago Press.

Lehmann, C. (1985). Grammaticalization: Synchronic variation and diachronic change. Lingua e Stile XX: 303-318.

Levshina, N. (2015). How to do linguistics with R: Data exploration and statistical analysis. Amsterdam: John Benjamins Publishing.

Traugott, E. C., & Trousdale, G. (Eds.). (2007). Gradience, Gradualness and Grammaticalization. Amsterdam: John Benjamins Publishing.

Combinatory principles of Chinese event nouns

Shan Wang

University of Macau

The selective nature of linguistic expression reflects a strong focus on describing social activities, revealing the inherent sociality and subjectivity of language use. However, systematic studies on the combinatory principles of Chinese event nouns with different word types remain limited. This study investigates the combination patterns and semantic preferences of event nouns, examining the distribution and usage tendencies of natural and non-natural event nouns when paired with various word types. By conducting a large-scale corpus analysis, this research reveals the underlying characteristics of different event nouns. The findings provide new empirical evidence on the combinatory principles of event nouns, highlighting a significant tendency toward non-natural event nouns in linguistic expression. Ultimately, the results indicate the social nature of language and the semantic interdependencies inherent in its use.

Navigating the complexities of aviation maintenance documentation: A corpus-based diachronic analysis of Simplified Technical English

Wanwen Amber Wang, Eric Friginal

Department of English and Communication, Hong Kong Polytechnic University

In aviation, where adherence to maintenance instructions is critical, unclear technical documentation can have lethal consequences. The prevalence of complex syntax and ambiguous terminology presents significant safety risks, particularly as a large portion of the global workforce comprises non-native English speakers (Friginal et al., 2020). Simplified Technical English (STE) has served as an international specification since its initial guide release in 1986, addressing these challenges and evolving through continuous user feedback. However, its effectiveness in reducing text complexity remains an underexplored area in current research.

This study investigates a diachronic corpus of Boeing 737 maintenance manuals (8.2 million tokens) spanning from pre-1990 to 2024, comparing Classic and Next Generation variants. Employing Biber's Multi-Dimensional Analysis (1988) alongside the computational tool Coh-Metrix (Graesser et al., 2004) to analyze text cohesion and complexity, the study uncovers significant linguistic variations in both informational production (Dimension 1) and online information elaboration (Dimension 6), highlighting the systematic evolution of aviation documentation practices. Follow-up experiments isolating eight distinctive linguistic features identified within these dimensions yield two key findings: (1) Next Generation manuals have transitioned from Classic formal technical writing to a more accessible plain language style, and (2) a trade-off has emerged regarding controlled linguistic features, where improvements in syntactic simplicity and temporality come at the cost of textual cohesion. These findings illustrate that linguistic simplification can be a double-edged sword; while it aims to enhance accessibility, advances in one dimension may inadvertently diminish essential textual attributes, such as natural language patterns and cohesive elements that facilitate comprehension. This corpus-based research offers valuable insights for the future development of STE, emphasizing the need to balance the competing demands of immediate comprehensibility and operational safety with the preservation of textual cohesion and appropriate linguistic complexity.

List of references

Biber, D. (1988). Variation across Speech and Writing. Cambridge University Press.

Friginal, E., Mathews, E. & Roberts, J. (2020). English in Global Aviation: Context, Research, and Pedagogy. Bloomsbury Publishing.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. Behavior Research Methods, Instruments & Computers, 36(2), 193–202. https://doi.org/10.3758/BF03195564

The impact of extramural English activities on phraseological complexity in L2 writing

Ying Wang[1], Henrik Kaatari[2], Tove Larsson[3], Taehyeong Kim[3], Pia Sundqvist[4]

[1]Karlstad University; [2]University of Gävle; [3]Northern Arizona University; [4]University of Oslo

Phraseological expressions, i.e., multi-word expressions including collocations (e.g., *heavy rain, make a decision*), play an important role in language acquisition and use (Wray, 2012). While L1 users acquire these expressions through extensive input, L2 learners, who traditionally rely more on formal instruction, often struggle with phraseology (Granger, 2019). However, in many countries, L2 learners are increasingly exposed to English outside the classroom through self-initiated Extramural English (EE) activities (Sundqvist, 2009). This study investigates the following research questions:

1) To what extent are there differences based on learners' engagement (vs. non-engagement) with five EE activities in terms of phraseological complexity of adjective-noun and verb-noun combinations?

2) How much does the impact of time spent differ across the activities?

The data were extracted from the Swedish Learner English Corpus (SLEC), which comprises argumentative texts written by Swedish secondary school students (years 7–11) and provides information on time spent on five EE activities: reading, watching, conversing, gaming, and social media (Kaatari et al., 2024). Following Vandeweerd et al. (2023), phraseological diversity was measured using moving windows. Given issues identified with commonly-used techniques to assess sophistication (e.g., register mismatches when calculating pointwise mutual information (PMI) from a reference corpus, Paquot & Hubert, 2025), we relied on comparisons within learner data to assess L2 phraseological sophistication.

The results indicate that the effectiveness of EE activities varies based on the *type* of EE activity and the *type*of phraseological units. Specifically, reading has a positive impact on adjective-noun combinations, while gaming and conversation show distinct patterns related to the amount of time students spend on these activities. These findings highlight the context-dependent nature of phraseology learning and suggest that classroom instruction could benefit from incorporating activities that engage learners in both input and outputand balancing creativity and conventionality to enhance phraseological development.

List of references

Granger, S. (2019). Formulaic sequences in learner corpora: Collocations and lexical bundles. In Siyanova-Chanturia, A., & Pellicer-Sanchez, A. (eds.) Understanding formulaic language: A second language acquisition perspective, pp. 228–247. Routledge.

Kaatari, H., Wang, Y., & Larsson, T. (2024). Introducing the Swedish Learner English Corpus: a corpus that enables investigations of the impact of extramural activities on L2 writing. Corpora, 19(1), 17–30.

Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. Second Language Research, 35, 121–145.

Paquot, M., & Hubert, N. (2025). Phraseological sophistication as a multidimensional construct: Exploring the relationship between association, register specificity and frequency of word combinations. International Journal of Learner Corpus Research, 11(1), 217–244.

Sundqvist, P. (2009). Extramural English matters: Out-of-school English and its impact on Swedish ninth graders' oral proficiency and vocabulary (Karlstad University Studies, 2009:55) [Doctoral dissertation, Karlstad University]. DiVA. https://www.diva-portal.org/smash/get/diva2:275141/FULLTEXT03.pdf

Vandeweerd, N., Housen, A., & Paquot, M. (2023). Comparing the longitudinal development of phraseological complexity across oral and written tasks. Studies in Second Language Acquisition, 45, 787–811.

Wray, A. (2012). What do we (think we) know about formulaic language? An evaluation of the current state of play. Annual Review of Applied Linguistics 32: 231–254.

Lexical complexity development of Chinese EFL learners in two genres and its relationship with writing quality

Zhihong Wang, Lianrui Yang

College of Foreign Languages, Ocean University of China

This study investigates the development of lexical complexity (LC) in Chinese EFL learners' writing across different proficiency levels and genres. While lexical knowledge plays a significant role in language proficiency, research focusing exclusively on LC development remains underrepresented compared to studies on syntactic complexity. We aim to build upon and extend the work of Yoon & Polio (2017) and Crossley & Macnamara (2015), who examined these effects in adult L2 English learners using holistic indices of syntactic complexity. Our research addresses a critical gap in understanding how LC develops across different genres and proficiency levels, particularly among Chinese EFL learners. We explore three key areas: LC development over time, genre influence, and the predictive power of lexical measures on writing quality.

Using a longitudinal design, we collected writing samples from 65 Chinese EFL learners spanning beginning, intermediate, and high proficiency levels. Participants completed 3 narrative and 3 argumentative writing tasks in turn throughout one semester. We analyzed LC using multiple measures including lexical diversity, sophistication, and density, and employed random effects modeling to examine both individual and group-level developmental patterns.

Our results revealed significant individual variation in lexical development trajectories, especially at beginning levels, with more consistent patterns emerging at higher proficiency levels. The findings reveal genre effects play a significant role in shaping lexical complexity, indicating that different text types elicit varying levels of lexical sophistication from learners. Notably, among all lexical features examined, lexical diversity emerged as the strongest predictor of writing quality, demonstrating the highest correlation with overall writing scores at beginner level.

Our findings suggest the need for dynamic, individualized approaches to L2 writing instruction and assessment, particularly for lower proficiency learners. This research enhances our understanding of lexical complexity development across proficiency levels and genres in Chinese EFL contexts.

Discourse analysis of public submissions to Australian GM reviews: A genre perspective

Susan Whitbread

Monash University

Controversy over genetic modification (GM) in food and agriculture is an issue that people take sides on. My research investigated and compared the linguistic choices made on this topic by four key stakeholder groups (general public, science, agribusiness, and government/regulatory) using data from public submissions to six government reviews into GM regulation in Australia, from 2016–2019. Submissions were downloaded, cleaned and used to create a specialised corpus of more than 1 million tokens in Sketch Engine. After an initial descriptive analysis of the corpus, I undertook a genre analysis to investigate this largely unreported type of public policy data.

In a genre analysis, communicative intent and context are important preliminary considerations (Fairclough, 1992; Bhatia, 2014). For context, I considered the role that public submissions play in a broader government change process, and concluded that the communicative intent of the submissions was persuasion. I then manually analysed the key 'moves' in my data and identified five (such as how the respondent demonstrated their credentials) that could help typify—and distinguish—the genre. Subsequently, I identified several trigger terms (e.g. *should, not and risk*) that could reasonably contribute to the persuasive intent of the submissions (Halmari & Virtanen, 2005), and subjected these to corpus concordance analyses, comparing and contrasting the four stakeholder groups. Finally, I considered the role of digital activist platforms in facilitating submissions to the reviews: the emergence of this form of facilitated submission raises questions about how such submissions conform to genre expectations, or whether new or 'genre-defiant' forms (Basgier, 2017) could be considered as part of the genre.

Overall, I concluded that public submissions were a novel genre, characterised by a communicative purpose of persuasion, within a public policy change context. This research has implications for how governments consider the greatly increasing levels of data obtained from public submissions processes.

List of references

Basgier, C. (2017). Atypical rhetorical actions: Defying genre expectations on Amazon.com. In C. R. Miller and A. R. Kelly (Eds.), Emerging genres in new media environments (pp. 187–202). Palgrave MacMillan. https://doi.org/10.1007/978-3-319-40295-6_10

Bhatia, V. (2014). Worlds of written discourse. A genre-based view. Bloomsbury Classics. ISBN: 978-1-4725-2263-4

Fairclough, N. (1992). Discourse and text: Linguistic and intertextual analysis within discourse analysis. Discourse and Society, 3(2), 193–217. www.jstor.org/stable/42887786

Halmari, H. & Virtanen, T. (Eds.). (2005). Persuasion across genres: A linguistic approach. John Benjamins Publishing Company.

Beyond the health crisis: A lexical multidimensional analysis of Bolsonaro's pandemic livestreams

Mirella Whiteman, Arianne Brogini, Marcos Oliveira, Aline Zamboni Milanez

Pontifical Catholic University of São Paulo

During the global Covid-19 health crisis, the communication strategy of former Brazilian President Jair Bolsonaro's administration raised both criticism and praise, as it conveyed official statements on YouTube live streams. Adopting strategies which emphasize nationalism, patriotism, respect for law and order, and opposition to measures identified as leftist (van Dijk, 2024), the president encouraged a return to normal activities. Meanwhile, Brazil ranked among the countries with the highest Covid-19 death tolls (Taylor, 2022). Recognizing that this same style of communication was adopted by other leaders around the world, the current paper investigates how these discursive patterns emerge in political communication during crises. Specifically, it analyzes the discourses, ideologies, and other lexis-based constructs (Berber Sardinha, 2024) in 100 live broadcasts delivered by Bolsonaro between 2020 and 2022. Lexical Multidimensional Analysis (Berber Sardinha; Fitzsimmons-Doolan, 2025) was employed to determine the predominant discursive dimensions. Transcriptions were made by hand and then tagged with the TreeTagger software for linguistic categorization. The lexical data were normalized for comparability and subjected to factorial analysis using SAS software. Three discursive dimensions emerged, whose tentative labels are 1) Institutional Delegitimization and Conservative Resistance; 2) Technocratic Governance and Market-Driven Ideology; 3) Libertarianism and Anti-Statism. The discourses indexed by the dimensions signal an agenda based on deregulation, opposition to leftist regimes, promotion of national sovereignty, advocacy for individual freedoms, market liberalization, and criticism of redistributive and environmental policies, as well as of the electoral system. The findings suggest ideological discourses that downplayed the severity of the health crisis.

List of references

Berber Sardinha, T. Exploring Multimodal Corpora In The Classroom from a Multidimensional Perspective. In: Crosthwaite, P. (org.). Corpora for Language Learning: Bridging the Research-Practice Divide. New York: Routledge, 2024. p. 25-42.

Berber Sardinha, T.; Fitzsimmons-Doolan, S. Lexical Multidimensional Analysis. Cambridge: Cambridge University Press, 2025.

Taylor, L. BMJ : British Medical Journal (Online); London Vol. 377, 2022.

Van Dijk, T. A. Discourse and ideologies of the radical right. Cambridge: Cambridge University Press, 2024.

Identifying discourses within multimodal surveillant landscapes through automatic text extraction from urban signage: A corpus-assisted discourse analysis

Viola Wiegand

University of Stirling

Surveillant practices are widespread in modern life, and while often based on digital technologies, some can be visible in physical environments, especially in cities. Recent years have seen the development of a sociolinguistics of surveillance, including qualitative approaches such as the multimodal analysis of 'surveillant landscapes' (Jones, 2017). By taking a corpus-assisted discourse analysis approach to the signage of surveillant landscapes, this talk aims to contribute to the sociolinguistics of surveillance and corpus linguistic approaches to multimodal data, especially images (see e.g. Malamatidou, 2020). It addresses the research question "How are discourses of surveillant landscapes characterised through urban signage?". To capture textual patterns in the multimodal data, the talk introduces an innovative method of corpus compilation: photographs of urban signage indicating that surveillance is in operation are processed with automatic text extraction. The extracted text is uploaded to Sketch Engine (Kilgarriff et al., 2014; https://www.sketchengine.eu/) to identify textual evidence with a focus on frequent content words and n-grams. Results point to e.g. verb forms such as *monitored*, *installed*, and *recorded* frequently displayed on signs of surveillance, while common patterns including *CCTV in operation* and *images are being monitored* highlight the role of urban video surveillance. The textual patterns, in relation to the original images, are interpreted through concordances and according to the surveillant landscape framework. The study examines how the signage of surveillant landscapes a) announces surveillant practices, b) evokes relationships between passers-by and operators, and c) indicates the potential internalised psychological effects and/or how information is recorded. I argue that a corpus linguistic approach is particularly suited to identifying textual patterns across a range of surveillant – and linguistic – landscapes, and that automatic text extraction together with more qualitative analyses of images has the potential to inform the corpus-assisted discourse analysis of multimodal data.

List of references

Jones, R. H. (2017). Surveillant landscapes. Linguistic Landscape, 3(2), 150–187.

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.

Malamatidou, S. (2020). Multimodality II: Text and image. In S. Adolphs & D. Knight (Eds.), Routledge Handbook of English Language and Digital Humanities (pp. 85–106). Routledge.

Prompt Engineering for complex discursive phenomena

Ell Wilding, Matteo Fuoli

Univerisity of Birmingham

One of the problems facing corpus linguists is the time-consuming and costly process of manually annotating corpora. Recent studies suggest that Large Language Models (LLMs) offer a potential solution by enabling the automatic annotation of pragmatic and discourse-level features, such as apologies (Yu & Li & et al. 2024) and rhetorical moves (Yu & Bondi & et al. 2024), at near human-level accuracy. However, these studies focus on relatively simple and formulaic features and only test a limited set of prompt engineering strategies. In this study, we propose and evaluate a range of new strategies for LLM-assisted corpus annotation and apply them to a broader spectrum of discursive phenomena with varying levels of complexity. The study is conducted in two phases. In phase 1, we draw on the prompt engineering literature in NLP and best practice guidelines from LLM developers to design 38 variations of a prompt instructing the LLM to annotate a text. We compare the performance of these prompts on the task of annotating instances of apologies in the Enron Trader Tapes Corpus, a large dataset of authentic telephone conversations involving Enron employees recorded during the California electricity crisis (Fuoli et al. forthcoming). We evaluate each prompt on GPT-4, GPT-4o, and Claude 3.5 Sonnet LLMs, conducting three runs per model to assess the consistency of their outputs. Accuracy is measured by comparing the models' outputs against a gold-standard, manually annotated corpus. In the second stage, we test five additional moves from Fuoli et al.'s (forthcoming) discursive trust management framework, covering different levels of complexity. Using OpenAI's API, we apply the three most successful prompts from the pilot stage (using negative directives, repeating yourself, and using cues) to a sample of texts containing these moves. Preliminary results show an inverse relation between accuracy and complexity of the codes.

List of references

Fuoli, Matteo & Nix, Adam & Wickert, Alicia & Van Riper, Annina. forthcoming. Trust, discourse, and corporate corruption: The case of Enron. Cambridge University Press.

Yu, Danni & Bondi, Marina & Hyland, Ken. 2024. Can GPT-4 learn to analyse moves in research article abstracts? Applied Linguistics amae071.

Yu, Danni & Li, Luyang & Su, Hang & Fuoli, Matteo. 2024. Assessing the potential of LLM-assisted annotation for corpus-based pragmatics and discourse analysis: The case of apologies. International Journal of Corpus Linguistics.

Leveraging corpus linguistics to address word choice errors in upper-secondary L2 academic writing in English

Lynn Williams Leppich

Bern University of Teacher Education

Corpus linguistics (CL) offers upper-secondary language teachers valuable data-informed insights that can drive instruction, curriculum design and student success (Biber et al., 2019). This study focuses on using CL to identify common word interference errors in English as a Foreign Language (EFL) writing among students preparing for the Cambridge C1 exam. By pre-identifying these errors, the aim is to improve writing instruction and learner outcomes (Gilquin, 2013).

The research explores two main questions: What are the most common word choice interference errors in upper-secondary EFL writing, and how can CL tools be used to identify and address these errors in the classroom? To research these questions, the study uses representative academic texts by learners for corpus-informed teaching intervention. CL tools such as concordancers will help identify recurring word choice errors and shape instructional materials and input to help students correct them (Sheldon, 2017).

The methodology includes conducting concordance analysis to identify patterns of error and analyse their frequency across the texts (McEnery et al., 2015). Based on the errors identified, targeted instructional activities will be designed, such as vocabulary exercises, error correction tasks and peer feedback activities, all informed by corpus data.

The findings of this study are expected to have significant pedagogical implications. By pre-identifying word choice errors, teachers can offer more targeted, data-driven instruction that addresses student weaknesses and optimises exam preparation (Laufer & Girsai, 2017). Additionally, this corpus-informed approach can be used to track student progress and refine the curriculum over time (Nesi & Gardner, 2012).

Overall, this research will demonstrate how CL can help teachers identify common interference errors and develop effective instructional strategies. By focusing on specific word choice errors and incorporating real student data, teachers can improve their students' academic writing proficiency and better prepare them for high-stakes exams (Liu & Zhang, 2017).

List of references

Biber, D., Conrad, S., & Reppen, R. (2019). Corpus linguistics: Investigating language structure and use (2nd ed.). CUP.

Gilquin, G. (2013). Data-driven learning and corpus-based teaching materials: A study of error analysis in English as a foreign language. CUP.

Laufer, B., & Girsai, N. (2017). The role of L1 interference in second language acquisition. CUP.

Liu, D., & Zhang, Y. (2017). Corpora and language teaching: A multi-disciplinary perspective. Palgrave Macmillan.

McEnery, T., Xiao, R., & Tono, Y. (2015). Corpus-based language studies: An advanced resource book (2nd ed.). Routledge.

Nesi, H., & Gardner, S. (2012). A corpus-based study of academic writing in English. Routledge.

Sheldon, C. (2017). Applying corpus-based techniques in second language writing instruction. OUP.

Linguistic synesthesia and sensory language: A meta-analysis

Bodo Winter[1], **Francesca Strik-Lievers**[2]

[1]University of Birmingham; [2]University of Genova, Department of Modern Languages and Cultures

Linguistic synesthesias combine different senses, as in the English expressions “smooth melody” (touch→sound) or “sweet sound” (taste→sound). For nearly a century (Ullmann, 1937), researchers have analyzed data from literary and general corpora to support the idea of a hierarchical ordering of the senses. According to this proposal, expressions map the presumed-to-be “lower” senses of touch, taste, and smell onto the presumed-to-be “higher” senses of sound and sight. In this study, we conduct the first-ever meta-analysis of linguistic synesthesias, synthesizing 38 datasets from corpus and dictionary studies of14 different languages (Ancient Greek, Chinese, Japanese, Latin, Turkish, Tzotzil, German, Spanish, French, Italian, Korean, Romanian, Hungarian, English), with data spanning from 1937 to the present and including both literary and non-literary language.

Our analysis reveals that there are consistent patterns in the data that are stable across languages. Using multiple analytical approaches, we highlight how theoretical conclusions depend on methodological choices. For example, prior studies often reported a metric of “hierarchy congruency” (e.g., Kumcu, 2021; Shen, 1997; Strik-Lievers, 2015; Winter, 2019), which measures the proportion of expressions that align with the hierarchy (e.g., “rough sound”, “sweet melody”) versus those that do not (e.g., “squealing touch”, “melodious sweetness”). Applying this metric across all datasets and controlling for cross-linguistic variation using a multilevel model with a random effect for language, we find that approximately 90% of all cases appear to support the hierarchy. However, closer examination reveals that this high average is driven by a few dominant mappings, specifically: touch→sound, touch→sight, and touch→sound. Taken together, these three mappings alone account for two-thirds of all hierarchy-congruent cases. These results suggest that the hierarchy may be an artifact of aggregation that falls apart once analytical approaches are used that do not lump expressions into a binary divide of “congruent” and “incongruent” cases.

List of references

Kumcu, A. (2021). Linguistic Synesthesia in Turkish: A Corpus-based Study of Crossmodal Directionality. Metaphor and Symbol, 36(4), 241–255. https://doi.org/10.1080/10926488.2021.1921557

Shen, Y. (1997). Cognitive constraints on poetic figures. Cognitive Linguistics, 8(1), 33–72. https://doi.org/10.1515/cogl.1997.8.1.33

Strik-Lievers, F. (2015). Synaesthesia: A corpus-based study of cross-modal directionality. Functions of Language, 22(1), 69–95.

Ullmann, S. (1937). Synaesthetic metaphors in William Morris. (An essay on the decorative art of the pre-raphaelites). Angol Filológiai Tanulmányok/Hungarian Studies in English, 2, 143–151.

Winter, B. (2019). Sensory linguistics: Language, perception, and metaphor. John Benjamins.

From birds to words: Onomatopoeia and metaphor in a corpus of field guides

Bodo Winter, **Marcus Perlman**

University of Birmingham

Language enables people to describe their sensory perceptions in exquisite detail, yet some perceptions are more challenging to describe than others due to uneven lexical differentiation across sensory modalities (Levinson & Majid, 2014). For example, English has richer vocabulary for visual concepts than for touch, taste, smell, or sound (Winter et al., 2018). To bridge these gaps, speakers often use strategies like metaphor or onomatopoeia (Winter, 2019).

This study examines linguistic strategies for describing bird vocalizations in field guides, such as the following description for Blue Rock Thrush (*Monticola solatarius*): "Calls include a flutey 'tiu-tee' and a vibrant 'wee-tchuk-tchuk', the latter notes like two stones being knocked together." Having built a corpus of ~6,000 descriptions from 20 guides covering ~2,800 bird species globally, we find that onomatopoeia is one of the primary linguistic strategies used, appearing in 70% of all entries. In contrast, in similar corpus analyses of descriptions of musical instruments (Wallmark, 2019), as well as descriptions of classical music more generally, only 0-2% onomatopoeias have been identified (Pérez-Sobrino & Julich, 2014). By contrast, comparative descriptors (e.g., "flute-like" or "sounds like a screeching door") only characterized 15% of entries.

Our findings show that onomatopoeias are consistently marked typographically, consistent with the more general finding that iconic words tend to be structurally marked (Dingemanse & Akita, 2017; Winter et al., 2023). Further analyses show that nearly 74% of onomatopoeias are combined with sensory adjectives like "sharp", "crisp", "clear", "soft", or "silvery". These adjectives are synesthetic metaphors (Shen & Cohen, 1998; Strik-Lievers, 2015) and characterize the timbre of sound (Wallmark & Kendall, 2018). We propose that there is a division of labor between onomatopoeia and synesthetic metaphor: onomatopoeias capture the overall temporal structure and melody of a vocalization, whereas sensory adjectives express timbre, which otherwise cannot be rendered in written onomatopoeias.

List of references

Dingemanse, M., & Akita, K. (2017). An inverse relation between expressiveness and grammatical integration: On the morphosyntactic typology of ideophones, with special reference to Japanese. Journal of Linguistics, 53(3), 501–532. https://doi.org/10.1017/S002222671600030X

Dingemanse, M., & Thompson, B. (2020). Playful iconicity: Structural markedness underlies the relation between funniness and iconicity. Language and Cognition, 12(1), 203–224. https://doi.org/10.1017/langcog.2019.49

Levinson, S. C., & Majid, A. (2014). Differential ineffability and the senses. Mind & Language, 29(4), 407–427. https://doi.org/10.1111/mila.12057

Miller, G. A., & Johnson-Laird, P. N. (1976). Language and perception. Harvard University Press.

Pérez-Sobrino, P., & Julich, N. (2014). Let's talk music: A corpus-based account of musical motion. Metaphor and Symbol, 29(4), 298–315.

Shen, Y., & Cohen, M. (1998). How come silence is sweet but sweetness is not silent: A cognitive account of directionality in poetic synaesthesia. Language and Literature, 7(2), 123–140. https://doi.org/10.1177/096394709800700202

Strik-Lievers, F. (2015). Synaesthesia: A corpus-based study of cross-modal directionality. Functions of Language, 22(1), 69–95.

Wallmark, Z. (2019). A corpus analysis of timbre semantics in orchestration treatises. Psychology of Music, 47(4), 585–605. https://doi.org/10.1177/0305735618768102

Wallmark, Z., & Kendall, R. A. (2018). Describing sound: The cognitive linguistics of timbre. In E. I. Dolan & A. Rehding (Eds.), The Oxford Handbook of Timbre. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190637224.013.14

Winter, B. (2019). Sensory linguistics: Language, perception, and metaphor. John Benjamins.

Winter, B., Lupyan, G., Perry, L. K., Dingemanse, M., & Perlman, M. (2023). Iconicity ratings for 14,000+ English words. Behavior Research Methods. https://doi.org/10.3758/s13428-023-02112-6

Winter, B., Perlman, M., & Majid, A. (2018). Vision dominates in perceptual language: English sensory vocabulary is optimized for usage. Cognition, 179, 213–220.

The formulaicity and intertextuality of sentencing remarks in criminal trials in England and Wales

David Wright, Helen Newsome-Chandler, James Thornton

Nottingham Trent University

This paper uses corpus-assisted discourse analysis to bring new insights into the form(s) and function(s) of sentencing remarks created and delivered by Crown Court judges in England and Wales. If a defendant pleads or is found guilty of a criminal offence in the Crown Court, the judge must assess all aspects of the offence and offender to pass a sentence that is fair and proportionate. The Sentencing Act (2020) states that judges 'must state in open court, in ordinary language and in general terms, the court's reasons for deciding on the sentence'. The Act also stipulates that, in determining a sentence, judges must follow any sentencing guidelines that are relevant to the case. Sentencing Guidelines provide direction on factors such as the harm caused to the victim, how culpable the offender is, and any aggregating or mitigating factors that may affect the sentence (Sentencing Council 2025).

The little linguistic research that has been conducted on sentencing remarks in England and Wales has focused on a relatively small sample of murder trials and has examined judges' appraisal and evaluation of people and actions (e.g. Potts and Weare 2018; Dai 2023). This paper describes the collection and analysis of a corpus of almost 400 sentencing remarks published between 2019 and 2024, covering an unprecedented range of offence types. The extraction of frequently occurring word n-grams reveals a high degree of formulaic, routinised expressions used by judges when passing sentences. These n-grams reflect judges' discursive strategies in fulfilling their obligations mandated by sentencing guidelines and are important manifestations of the intertextual networks that determine the form and content of remarks. These findings help us better understand the consistency and transparency of language in sentencing remarks, which are important factors in ensuring public understanding of sentencing (House of Commons Justice Committee 2023).

List of references

Dai, X. (2023) With or without a purpose? Judges' appraisal of offenders or their behaviour in six sentencing remarks. Text & Talk, 43(4), 449 – 469. https://doi.org/10.1515/text-2020-0228

House of Commons Justice Committee (2023) Public Opinion and Understanding of Sentencing Available at: https://committees.parliament.uk/publications/41844/documents/207521/default/ (Accessed: 16 January 2025)

Potts, A. and Weare, S. (2018) Mother, Monster, Mrs, I: A critical evaluation of gendered naming strategies in English sentencing remarks of women who kill. International Journal for the Semiotics of Law, 31(1), 21–52 https://doi.org/10.1007/s11196-017-9523-z

Sentencing Act 2020, c. 17. Available at: https://www.legislation.gov.uk/ukpga/2020/17/contentsm (Accessed: 16 January 2025).

Sentencing Council (2025) About sentencing guidelines. Available at: https://www.sentencingcouncil.org.uk/sentencing-and-the-council/about-sentencing-guidelines/ (Accessed: 16 January 2025)

Are new farmers really new in China?: A corpus-assisted discourse study

Zihan Xia[1], **Hao Huang**[2]

[1]Department of English and Communication,The Hong Kong Polytechnic University; [2]Department of Linguistics and Translation,City University of Hong Kong

The representation of farmers remains a pivotal issue, intersecting with a broad range of political concerns (Day, 2013). Given the potential influence of official media reports on shaping public knowledge and understanding of farmers (Mayr, 2015), a critical examination of their representation is warranted. This study investigates the discursive construction of farmers in Chinese official media, with a specific focus on People's Daily, one of the official organs representing Chinese government. A corpus of reports containing the word 农民 (farmer) in both title and body, spanning the period from 2007 to 2017, was compiled to achieve the purpose. The corpus contains 4,014 news reporting, with more than 3 million Chinese words (after segment). Notably, 2012 marks the midpoint of this period, aligning with the launch of the rural revitalisation initiative in China. This underscores the necessity to analyse how policy shifts have influenced media representation during this time frame. Employing a corpus-assisted discourse analysis approach (Baker et al., 2008) with consistent collocation analysis (Gabrielatos & Baker, 2008) and peaks and troughs analysis (Gabrielatos et al., 2012), this study examines the frames used to represent farmers in both synchronic and diachronic perspectives. The findings reveal a preponderance of reports focused on financial development, an upward trend in individual agency of farmers, and the emergence of a policy-oriented 'new' farmer representation, which simultaneously challenges and reinforces existing stereotypes. The policy focus of different periods is found to shape new discourses, influencing the reporting focus of official media and, in turn, constructing normative representations of farmers in each period.

List of references

Baker, P., Gabrielatos, C., KhosraviNik, M., Krzyżanowski, M., McEnery, T., & 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 & Society, 19(3), 273–306. https://doi.org/10.1177/0957926508088962

Day, A. F. (2013). The Peasant in Postsocialist China: History, Politics, and Capitalism. Cambridge University Press. https://doi.org/10.1017/CBO9781139626309

Gabrielatos, C., & Baker, P. (2008). Fleeing, Sneaking, Flooding: A Corpus Analysis of Discursive Constructions of Refugees and Asylum Seekers in the UK Press, 1996-2005. Journal of English Linguistics, 36(1), 5–38. https://doi.org/10.1177/0075424207311247

Gabrielatos, C., McEnery, T., Diggle, P. J., & Baker, P. (2012). The peaks and troughs of corpus-based contextual analysis. International Journal of Corpus Linguistics, 17(2), 151–175. https://doi.org/10.1075/ijcl.17.2.01gab

Mayr, A. (2015). Institutional Discourse. In D. Tannen, H. E. Hamilton, & D. Schiffrin (Eds.), The Handbook of Discourse Analysis (Vol. 2, pp. 755–774). Wiley-Blackwell.

Voice and engagement in museum exhibit labels: A comparative corpus-assisted discourse study of two art museums

Xiaoyu Xu

The Education University of Hong Kong

Museum exhibit labels have historically been characterised by a dominant 'objective' institutional-authority voice. However, in recent decades, there has been a noticeable paradigm shift towards democratisation and the inclusion of diverse perspectives. Despite this transformative trend, regions like Hong Kong, where the researcher is based, have seen limited applied linguistics research dedicated to developing a toolkit of linguistic resources for engaging voices in exhibit labels. This dearth of textual analysis poses challenges to the evolution of professional practices and impedes the training of museum writers.

This research attempts to explore textual resources that support the inclusion of varied voices in exhibit labels by comparing practices between a museum in Hong Kong and one in the UK. To this end, a corpus-assisted discourse study approach was adopted. In the summer of 2024, 150 labels from each museum were gathered and annotated using the engagement subsystem of the Appraisal framework developed by Martin and White (2005). Subsequent statistical analyses and meticulous data scrutiny were conducted using the UAM CorpusTool.

Contrasting approaches were found, with the Hong Kong museum adopting an authoritative tone characterised by monoglossic claims, while the UK museum embraced a more speculative and heteroglossic style, particularly evident in commentaries where open questions are often left to the visitors. Attributed claims, particularly from different artists, are particularly useful in adding dialogic dynamism. Moreover, the UK museum leaned towards a more disclaiming tone in introducing the exhibit title and artist bios to focus on how the current artist challenges previous practices, while the Hong Kong museum favoured a more assertive and proclaiming style to endorse the artist. This study sheds light on the nuances of label writing practices and how engagement devices construct voice dynamism. Useful resources can be transformed into training materials for fostering interactive and inclusive writing practices.

List of references

Blunden, J. (2017). The sweet spot? Writing for a reading age of 12. The Museum Journal, 60(3), 291-309.

Martin, J. R., & White, P. R. R. (2005). The language of evaluation: Appraisal in English. Basingstoke: Palgrave Macmillan.

Ravelli, L. (2006). Genre and the museum exhibition. Linguistic and the Human Sciences, 2(2), 299-317.

Xu, X. (2023, June 28-30). The role of language in visitor engagement in current Hong Kong virtual museums [paper presentation]. Joint AELFE-LSPPC International Conference, Zaragoza, Spain.

AI-Powered tools as academic English writing assistants to English-as-L2 Authors: A multi-dimensional analysis

Rogerio Yamada

Pontifical Catholic University of São Paulo

Non-English speakers are widely required to write in English, as the predominance of English in academia has established it as the primary medium for mainstream academic journals (Baumvol, Sarmento, & Da Luz Fontes, 2021; Belcher, 2007; Cargill & Burgess, 2008; Flowerdew, 2012). However, the challenge lies in writing idiomatic language that meets the register expectations of academic English. The academic register in English relies on specific rhetorical and lexicogrammatical patterns (Biber & Barbieri, 2007; Ädel & Erman, 2012) that differ from many of those in their authors' academic background (Pang, 2010; Wray, 2019; Ädel & Erman, 2012). Assuming that Large Language Models such as ChatGPT have been extensively trained with academic English language, this study aimed at verifying the degree to which AI-powered tools are capable of assisting English-as-L2 writers to meet the lexicogrammatical and rhetorical needs of academic English. Previous research in AI-generated academic English suggests that AI tools do not necessarily reproduce the rhetorical or lexicogrammatical choices of human authors (Berber Sardinha, 2024). To further explore this, the study employed an English-as-L2-Authored Papers (EL2AP) corpus compiled from the SciELO Preprints archive (SciELO Preprints, n.d.) and a Mainstream Journals Published Papers (MJPP) corpus reflecting the same areas of knowledge as EL2AP. Only articles submitted before the advent of ChatGPT were included. EL2AP texts were revised with ChatGPT and compiled into the AI-EL2AP corpus. To gauge the similarities and differences between the human and AI-generated texts, a Multi-Dimensional Analysis (Biber, 1988, 1995) was carried out through an additive analysis (Berber Sardinha, Pinto, & Biber, 2014). Overall, the results indicated that AI-assisted academic writing diverges from human standards by relying on non-typical patterns of academic English. The dimensional profiles of the AI-generated and human-authored texts will be detailed in the presentation.

List of references

Baumvol, L., Sarmento, S., & Da Luz Fontes, A. B. A. (2021, August). Scholarly publication of Brazilian researchers across disciplinary communities. Journal of English for Research Publication Purposes, 2 (1), 5–29. Retrieved 2024-10-18, from http://www.jbe-platform.com/content/journals/10.1075/jerpp.20012.bau doi: 10.1075/jerpp.20012.bau

Belcher, D. D. (2007, March). Seeking acceptance in an English-only research world. Journal of Second Language Writing , 16 (1), 1–22. Retrieved 2023-05-04, from https://linkinghub.elsevier.com/retrieve/pii/S1060374306000786 doi: 10.1016/j.jslw.2006.12.001

Berber Sardinha, T. (2024, April). AI-generated versus human-authored texts: A multidimensional comparison. Applied Corpus Linguistics, 4 (1), 100083. Retrieved 2024-01-05, fromhttps://linkinghub.elsevier.com/retrieve/pii/S2666799123000436 doi: 10.1016/j.acorp.2023.100083

Berber Sardinha, T., Pinto, M. V., & Biber, D. (Eds.). (2014). Multi-dimensional analysis, 25 years on: a tribute to Douglas Biber (No. volume 60). Amsterdam; Philadelphia: John Benjamins Publishing Company.

Biber, D. (1988). Variation across speech and writing (1st ed.). Cambridge: Cambridge University Press.

Biber, D. (1995). Dimensions of register variation: a cross-linguistic comparison. Cambridge; New York: Cambridge University Press.

Biber, D., & Barbieri, F. (2007, January). Lexical bundles in university spoken and written registers. English for Specific Purposes, 26 (3), 263–286. Retrieved 2024-12-04, from https://linkinghub.elsevier.com/retrieve/pii/S0889490606000366 doi: 10.1016/j.esp.2006.08.003

Cargill, M., & Burgess, S. (2008, April). Introduction to the Special Issue: English for Research Publication Purposes. Journal of English for Academic Purposes, 7 (2), 75–76. Retrieved 2024-10-18, from https://linkinghub.elsevier.com/retrieve/pii/S147515850800009X doi:10.1016/j.jeap.2008.02.006

Flowerdew, J. (2012, November). English for Research Publication Purposes. In B. Paltridge & S. Starfield (Eds.), The Handbook of English for Specific Purposes (1st ed., pp. 301–321). Wiley. Retrieved 2024-10-23, from https://onlinelibrary.wiley.com/doi/10.1002/9781118339855.ch16 doi: 10.1002/9781118339855.ch16

Pang, W. (2010, October). Lexical Bundles and the Construction of an Academic Voice: A Pedagogical Perspective. The Asian EFL Journal , 47 . Retrieved from https://asian-efl-journal.com/monthly-editions-new/lexical-bundles-and-the-construction-of-an-academic-voice-a-pedagogical-perspective/index.htm

SciELO Preprints. (n.d.). Retrieved 2025-01-14, from https://preprints.scielo.org/index.php/scielo

Wray, A. (2019). Concluding question: Why don't second language learners more proactively target formulaic sequences? In A. Siyanova-Chanturia & A. Pellicer-Sánchez (Eds.), Understanding formulaic language: a second language acquisition perspective (pp. 248–269). New York London: Routledge, Taylor & Francis Group. doi: 10.4324/9781315206615

Ädel, A., & Erman, B. (2012, April). Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. English for Specific Purposes, 31 (2), 81–92. Retrieved 2024-12-04, from https://linkinghub.elsevier.com/retrieve/pii/S0889490611000573 doi: 10.1016/j.esp.2011.08.004

A corpus-based textual cohesive analysis of translational language generated by machine

Ruiying Yang[1], **Liang Xu**[2]

[1]Xi'an Jiaotong University; [2]Jiangxi Agricultural University

With the continuous advancement of artificial intelligence technology, machine translation has become increasingly important, motivating the emergence of a large number of machine translation systems in recent years. Although, it is widely believed that the quality of current machine translation software DeepL is one of the highest in the world, scientific assessment of it is still rare. This study sets off to investigate the linguistic features and performance of DeepL in translating research article abstracts through the approach of quantitative linguistics. The analysis is based on 120 human-translated English abstracts from core agricultural journals in Chinese, their corresponding machine-translated English versions generated by DeepL, and 120 expert-authored English abstracts in international peer-reviewed journals in the same discipline. In order to find out whether there are distinctive differences in linguistic features among the above three types of texts, a total of 44 quantitative linguistic features were extracted using TAACO 2.0 (the Tool for the Automatic Analysis of Cohesion) to compare machine translation with human translation and expert-authored texts. The study found that machine-translated texts exhibit significant differences compared to human translations on 22 indices and to expert-authored texts on 33 indices. Specifically, machine translated texts demonstrate lower lexical density and higher lexical overlap than the other two groups. Machine translation tends to adopt a relatively simplistic and conservative approach to vocabulary, relying on repetitive words. At the syntactic level, machine-translated texts exhibit higher similarity between adjacent sentences, reflecting reduced text complexity but enhanced coherence. At the discourse level, machine translation frequently uses conjunctions to enhance textual cohesion. The distinctive linguistic characteristics revealed in this study underscores the utility of quantitative linguistic methods in evaluating the performance of machine translation systems.

List of references

Aslerasouli, P., & Abbasian, G. R. (2015). Comparison of google online translation and human translation with regard to soft vs. hard science texts. Journal of Applied Linguistics and Language Research, 2(3), 169-184.

Azer, H. S., & Aghayi, M. B. (2015). An evaluation of output quality of machine translation (Padideh Software vs. Google Translate). Advances in Language and Literary Studies, 6(4), 226-237.

Crossley, S. A., Kyle, K., & Dascalu, M. (2019). The Tool for the Automatic Analysis of Cohesion 2.0: Integrating semantic similarity and text overlap.Behavior Research Methods,51(1), 14-27.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. Journal of Second Language Writing, 32, 1-16.

Jiang, Y. (2014). A Quantitative Linguistic Comparison of Human Translations and Automatic Online Translations in Terms of Stylistic Features. Foreign Language Education, 35(5), 98-102.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). Automated evaluation of text and discourse with Coh-Metrix. Cambridge University Press.

Nasution, D. K. (2022). Machine Translation in Website Localization: Assessing its Translation Quality for Language Learning. Al-Ishlah: Jurnal Pendidikan, 14(2), 1879-1886.

Wang, P. (2023). A Comparative Study of Linguistic Quantitative Features in Human and Google Translations of Fiction and Social Science Texts (Master's thesis, Beijing Foreign Studies University).

Integrating pattern grammar and local grammar into the identification of constructions and the development of a pedagogical constructicon in EAP context: An exploratory study

Jun Ye

Sichuan International Studies University

This study explores the possibility of integrating pattern grammar and local grammar into construction grammar research and the value of such research for EAP writing pedagogy. Using data taken from a corpus compiled from Linguistics research articles, our study shows that local grammar analyses of grammar patterns help to identify 'pattern-meaning' combinations. We further describe the functions of such combinations in academic contexts and argue that they can be interpreted as constructions at the mid-level of specificity. We also demonstrate the organisation of these constructions into a constructicon for pedagogical purposes. The implications and applications are subsequently discussed, highlighting in particular that the constructicon not only offers an inventory of pattern-meaning pairings available to express meanings, but also provides quantitative information about the association between verbs and constructions. Overall, our study indicates that the integration of pattern grammar, local grammar and construction grammar could be of immense potential value for EAP writing research and pedagogy.

Wordless: An integrated corpus tool with multilingual support for the study of language, literature, and translation

Lei Ye

Shanghai International Studies University

This paper presents *Wordless* (Ye, 2024), an integrated corpus tool with multilingual support for the study of language, literature, and translation. It is a free, cross-platform, and open-source desktop application with a user-friendly graphical interface which is specially designed to cater the needs of non-technical users. Its ultimate goal is to eliminate all unnecessary technological barriers to the utilization of cutting-edge technologies by researchers in the field of corpus-based studies.

Wordless is "battery-included": it is integrated with built-in multilingual support for 120+ languages across the globe with the capability to conduct multiple NLP tasks including sentence/word/syllable tokenization, part-of-speech tagging, lemmatization, stop word filtering, dependency parsing, sentiment analysis. Configuration details of all data files and language models are meticulously handled by *Wordless*.

Wordless has implemented a wide variety of statistical measures including 39 readability formulas, 28 indicators of lexical density and diversity, 9 measures of dispersion, 8 measures of adjusted frequency, 8 tests of statistical significance, 2 measures of Bayes factor, and 18 measures of effect size. *Wordless* also comes with a large number of data visualization options including dispersion plots, line charts, word clouds, network graphs, and dependency graphs.

As of the writing of this paper, around 50 journal articles, MA and PhD theses, and conference papers have been completed with the aid of *Wordless*. Relevant literature show that *Wordless* has been successfully applied to various research areas including but not limited to cognitive linguistics (Sun et al., 2023), communication studies (Zhang & Hou, 2024), critical discourse analysis (Xu & Tao, 2023), language teaching (Ma, 2023), pedagogy (Lin, 2023), quantitative linguistics (Dai & Liu, 2024), L2 acquisition (Yi & DeKeyser, 2022), software engineering (Wei et al., 2023), and translation studies (Fu & Liu, 2024).

List of references

[1]Dai, Z., & Liu, H. (2024). Part-of-speech features in Bob Dylan's song lyrics: A stylometric analysis. International Journal of Humanities and Arts Computing, 18(2), 249–264. https://doi.org/10.3366/ijhac.2024.0335

[2]Fu, L., & Liu, L. (2024). What are the differences? A comparative study of generative artificial intelligence translation and human translation of scientific texts. Humanities and Social Sciences Communications, 11, Article 1236. https://doi.org/10.1057/s41599-024-03726-7

[3]Lin, Y. (2023). Lishi shuju keshihua zai gaozhong lishi jiaoxue zhong de yingyong yanjiu [Research on the application of historical data visualization in senior high school history teaching] [Master's thesis, Fujian Normal University]. CNKI. https://link.cnki.net/doi/10.27019/d.cnki.gfjsu.2023.002670

[4]Ma, L. (2023). Kōkō·daigaku ni okeru Nihongo kyōkasho no renketsu ni kansuru kenkyū: "Futsū Kōkō Kyōkasho Nigo" to "Shinseki Daigaku Nigo" o rei ni [A study on the articulation between the high school and college Japanese textbooks: A case study of High School Japanese and New Century Japanese] [Master's thesis, Harbin Normal University]. CNKI. https://link.cnki.net/doi/10.27064/d.cnki.ghasu.2023.001925

[5]Sun, Y., Kong, D., & Zhou, C. (2023). Economy or ecology: Metaphor use over time in China's government work reports. Language and Cognition, 15(3), 551–573. https://doi.org/10.1017/langcog.2023.18

[6]Wei, J., Chen, X., Xiao, H., Tang, S., Xie, X., & Li, Z. (2023). Natural language processing-based requirements modeling: A case study on problem frames. In Proceedings of 2023 30th Asia-Pacific Software Engineering Conference (APSEC) (pp. 191–200). IEEE. https://doi.org/10.1109/APSEC60848.2023.00029

[7]Xu, B., & Tao, Y. (2023). National identity in media discourses from Russia and Ukraine: Amid the 2022 Russo-Ukranian War. Zeitschrift für Slawistik, 68(3), 419–439. https://doi.org/10.1515/slaw-2023-0021

[8]Ye, L. (2024). Wordless: An integrated corpus tool with multilingual support for the study of language, literature, and translation. SoftwareX, 28, Article 101931. https://doi.org/10.1016/j.softx.2024.101931
[9]Yi, W., & DeKeyser, R. (2022). Incidental learning of semantically transparent and opaque Chinese compounds from reading: An eye-tracking approach. System, 107, Article 102825. https://doi.org/10.1016/j.system.2022.102825
[10]Zhang, H., & Hou, Y. (2024). The construction of interpersonal meanings in Jiaqi Li's e-commerce live streams: Integrating verbal and visual semiotics. Journal of Business and Technical Communication, 38(4), 371–409. https://doi.org/10.1177/10506519241258445

Family and marriage in a Japanese newspaper: Shifting representations over two decades

Keisuke Yoshimoto

Ryukoku University

This paper examines how a liberal Japanese newspaper has challenged traditional family ideologies and tracks shifts in family and marriage representations over the past two decades. A gap exists between the conservative government's focus on preserving patriarchal families and the reality of diverse family forms in Japan. Influenced by Confucian values, adult children are expected to marry, have children, and care for ageing parents. The importance of family units is reflected in the *koseki* system, or the family registry system, which requires family members to share the surname of the household head. However, Japan's population has been declining with the elderly (65+) making up about 30% of the population and children (0-14) just 11.3%. The birth rate reached a record low of 1.20 in 2023, reflecting rising rates of unmarried individuals. Societal pressures related to gender roles and family labour divisions contribute to these trends (Gender Equality Bureau Cabinet Office, 2024).

A corpus of 3.5 million words from *Mainichi* newspaper articles (2005-2024) focused on family (家族) and marriage (婚) was analysed. Following Brookes and Baker's (2021) 'remainder method', the corpus was divided into four subcorpora by five-year periods, then compared with the remainder of the corpus to generate keywords. The keywords were categorised by themes, showing how family-related representation has evolved. The study reveals a shift in women's frustrations, initially focused on post-divorce gender inequality, later transitioning to the issue of family name changes upon marriage. Since 2015, women's concerns have been framed alongside those of same-sex couples due to the introduction of same-sex partnership certificates. The focus on women's rights waned during the 2010-2014 period, marked by the Great Kanto Earthquake and the Fukushima disaster, which emphasised family formation over rights issues. However, this does not apply to the COVID-19 pandemic period, where family bonding and rights issues were discussed.

List of references

Brookes, Gavin & Baker, Paul. 2021. Obesity in the News: Language and Representation in the Press. Cambridge University Press.

Gender Equality Bureau Cabinet Office. 2024. The White Paper on Gender Equality 2024.

"Liberate women from the burden of family": The representation of feminism 女权 and related issues in The People's Daily

Zijin You

University of Glasgow

Feminism has been a notoriously contentious issue in China (Min, 2007; Spakowsi, 2011), with authorities metaphorizing feminism as a tumour (BBC, 2022; Yan, 2022; Chen, 2022).

Previous studies on the coverage of feminism in China have focused on social media platforms (Han, 2018; Wang & Driscoll, 2018; Mao, 2020) whereas those in official discourse has been neglected. Therefore, this study examines how the term *女权* has been reported in the Chinese state-approved official discourse, namely the newspaper *The People's Daily*. To do this, a corpus of all articles which use the term *feminism* *女权* (i.e. the 'direct' corpus) were built. Acknowledging that contentious issues are lexicalised in different ways, reflecting different framings of the same issue, this study develops an innovative methodology to help locate references to feminism which do not invoke the exact term *女权* and built the 'indirect' corpus. A diachronic analysis was then conducted. Three years have been selected for analysis. They were 1949 and 1995 (which were the years when the term *feminism* *女权* were most frequent) and 2022 as the most recent year.

Interestingly, this study notices that language representing the relationship between women and family has changed drastically overtime. Family was repeatedly framed as "a burden on women" in 1949 (in 16 out of 81 hits). In contrast, in 1995, only one example described family 家庭 as a "*shackle* *桎梏*" out of all 351 hits. In 2022, a conflict between women and family was not identified. It is unlikely that these dramatic changes in the representation of women's relationship to the family are random and this paper sets out the reasons for these linguistic phenomena.

List of references

BBC. (2022, April 14). China's Communist Youth League says "extreme feminism has become a cancer on the Internet", sparking controversy. https://www.bbc.com/zhongwen/trad/chinese-news-61105757

Chen, S. (2022, May 4). China's crackdown on online feminist activism. FairPlanet. https://www.fairplanet.org/editors-pick/chinas-crackdown-on-online-feminist-activism/

Han, X. (2018). Searching for an online space for feminism? the Chinese feminist group Gender Watch Women's Voice and its changing approaches to online misogyny. Feminist Media Studies, 18(4), 734–749.

Min, D. (2007). Duihua (dialogue) in-between. Interventions, 9(2), 174–193. https://doi.org/10.1080/13698010701409111

Mao, C. (2020). Feminist activism via social media in China. Asian Journal of Women's Studies, 26(2), 245–258.

Spakowski, N. (2011). "Gender" trouble: Feminism in China under the impact of western theory and the spatialization of identity. positions: east Asia cultures critique, 19(1), 31-54.

Wang, B., & Driscoll, C. (2018). Chinese feminists on social media: Articulating different voices, building strategic alliances. Continuum, 33(1), 1–15.

Yan, A. (2022, April 15). China Communist Youth League lashes out at "malignant tumour" feminism. South China Morning Post. https://www.scmp.com/news/people-culture/gender-diversity/article/3174419/china-communist-youth-league-lashes-out

Placing vulnerability and resilience in the climate adaptation debate: A corpus-based framing analysis of expert discourse in Australia

Lorenzo Zannini

University School for Advanced Studies IUSS Pavia/University of Naples "L'Orientale"

Vulnerability and resilience are two fundamental concepts within the climate adaptation debate. Mainstream conceptualisations of vulnerability and resilience reflect a depoliticised understanding of these concepts as stable attributes. Crucially, this approach fails to address the complex and dynamic nature of socio-ecological systems, resulting in a process of individualising responsibilities for coping with climate risks that obscures power differentials with critical implications for climate (in)action (Taylor, 2015). Studies, such as Bankoff (2019), have highlighted the need to discuss these concepts in their procedural character to foreground their contentious and power-laden nature rooted in historical injustices. While the conflict-ridden processes of intergovernmental negotiations and reporting have so far failed to translate climate adaptation knowledge into actionable discourses effectively (Bureau, 2024), looking at how experts directly communicate adaptation to the public might provide a way to access narratives that remain backgrounded in other discursive contexts. For this reason, the paper presents the results of a corpus-based framing analysis of vulnerability and resilience in a specialised corpus of 187 research-based news articles published between 2011 and 2023 in *The Conversation Australia*. Since knowledge production is a situated process embedded in power relations, the national focus on Australia will allow for an interrogation of how “culturally-mediated epistemologies” (Curry, 2024, p. 236) guide the social construction of knowledge in a context shaped by emerging neoliberal governmentality (Jackson, 2024), coloniality, and individualised responsibilities for coping with climate hazards. In so doing, the paper will address the role of expert discourse in delivering science to stakeholders to foster climate action. As part of a broader research project on adaptation discourse, this work explores how studies of under-researched domains such as research-based news can enhance the understanding of the discursive construction of climate-related concepts and phenomena and possibly enhance climate action.

List of references

Bankoff, G. (2019). Remaking the world in our own image: Vulnerability, resilience and adaptation as historical discourses. Disasters, 43(2), 221–239. https://doi.org/10.1111/disa.12312

Bureau, P. (2024). Climate knowledge or climate debate?: Using word embeddings and Critical Discourse Analysis to compare expert and media representations of climate knowledge. Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication, 30(1), 35–57. https://doi.org/10.1075/term.00076.bur

Curry, N. (2024). Questioning the climate crisis: A contrastive analysis of parascientific discourses. Nordic Journal of English Studies, 23(2), 235-267. https://doi.org/10.35360/njes.v23i2.39190

Jackson, G. (2024). The influence of climate resilience governmentality on vulnerability in regional Australia. Environment and Planning E: Nature and Space, 7(3), 1098-1121. https://doi.org/10.1177/25148486241226919

Taylor, M. (2015). The political ecology of climate change adaptation: Livelihoods, agrarian change and the conflicts of development. Routledge.

Exploring functional variability in native and non-native Czech texts

Adrian Jan Zasina

Charles University

The range of differences between native and non-native written production can be varied, but there is consensus on the presence of non-nativeness in learners' language. Numerous studies have affirmed that the language proficiency of adult learners differs from the system developed by native speakers (Clahsen & Felser, 2006). One approach to capturing differences between native and non-native texts may involve examining their functional variability through multidimensional analysis (Biber, 1989). Previous research (Staples et al., 2022; Weigle & Friginal, 2015), conducted in the English language, has demonstrated that the use of a multidimensional analysis makes it possible to uncover diverse characteristics of native and non-native writing. It also underscores that comparisons should be made within the same genres, as the topic of the text has a greater influence on variability than non/nativeness.

As of now, there is no research comparing the functional variability of Czech native and non-native texts. Hence, the aim of this paper is to expand our understanding of dissimilarities between native and non-native Czech by utilising a Czech model of multidimensional analysis (Cvrček et al., 2018). Two different datasets were employed: learner texts from Polish speakers at level A2 (62 texts from 16 students) and texts from Czech native speakers (32 texts from 8 students). All texts were written on four topics with the same instructions, encompassing an informal letter, a description of a place, an argumentative essay, and storytelling. Subsequently, both sets were projected onto the multidimensional space to facilitate a comparison.

Taken together, these findings suggest that non-native texts are generally more dynamic, employing verbal features more frequently. Nevertheless, there are similarities in more conventional genres, where both native and non-native argumentative essays focus on using nominal phrases. However, native argumentation reveals more complex nominal constructions and a higher degree of cohesion.

List of references

Biber, D. (1989). A typology of English texts. Linguistics, 27(1), 3–43. https://doi.org/10.1515/ling.1989.27.1.3

Clahsen, H., & Felser, C. (2006). How native-like is non-native language processing? TRENDS in Cognitive Sciences, 10(12), 564–570. https://doi.org/10.1016/j.tics.2006.10.002

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

Staples, S., Gray, B., Biber, D., & Egbert, J. (2022). Writing Trajectories of Grammatical Complexity at the University: Comparing L1 and L2 English Writers in BAWE. Applied Linguistics, 46–71. https://doi.org/10.1093/applin/amac047

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

Resilience of Taiwan representation in Czech Parliamentary discourse (2017–2021): A corpus-based study

Adrian Jan Zasina, Adam Horálek, Andrea Hudousková, Svatava Škodová

Charles University

The digitalization has led to an unprecedented volume of data, enabling the analysis of linguistic trends and public discourse. This study examines the representation of Taiwan and its resilience in Czech Parliamentary discourse from 2017 to 2021, focusing on how language use, occurrence and frequency reflects socio-political narratives. Two questions were posed:

1. How has the frequency of Taiwan-related lexemes (Taiwan, Taiwanese) evolved in Czech Parliamentary discourse over time?

2. What collocational patterns reveal semantic and pragmatic aspects of Taiwan's representation in Czech Parliamentary discourse and its resilience?

The research is based on the Corpus of Czech Parliamentary Speeches (Berrocal & Berrocal, 2021) that delivers speeches in the Lower Chamber of the Czech Parliament. Its core comprises electronic transcripts of parliamentary debates, which are publicly available at www.psp.cz. These texts include remarks made in response to other speakers, as well as comments by the Chair of the session. The analysis makes uses of time profiles to analyse the frequency of lexemes over time, while collocational profiles examine co-occurring words to uncover framing patterns. Consequently, anthropological method of thick description (Gertz, 2000) was applied to linguistic data for qualitative interpretation.

The time analysis reveals resilience and fluctuations in parliamentary attention to Taiwan that are depended on key socio-political events such as diplomatic visits, trade agreements, and discussions of international relations. Collocational patterns provide insight into contrasting representations of Taiwan in Czech parliamentary discourse, with some frames highlighting its technological and democratic identity, while others focus on contentious issues, such as geopolitical tensions with relation to China.

The findings provide valuable insights into the linguistic strategies used to frame Taiwan in Czech parliamentary discourse, offering a comparative perspective on political language and narratives. This research highlights the importance of corpus-driven methods for examining the intersection of language, geoeconomics, and international relations.

List of references

Berrocal, M- & Berrocal, M. (2021). ParlCorp. Corpus of Czech Parliamentary Speeches. Ústav Českého národního korpusu FF UK, Praha. Available on-line: http://www.korpus.cz

Geertz, C. (2000). The interpretation of cultures, 3–30. New York: Basic Books.

The Chamber of Deputies – official website providing information about the lower house of parliament. Available at www.psp.cz

From prompts to practice: A corpus-assisted study of AI-Generated texts for inclusive and sustainable tourism

Federico Zaupa, Silvia Cavalieri

University of Modena and Reggio Emilia

Today, one of the primary aims of using Artificial Intelligence (AI) has been the production of texts for special purposes. This phenomenon has also drawn, among others, the attention of discourse analysts employing corpus linguistics tools. Within this line of research, recent studies have focused on evaluating the features and quality of AI-generated texts in different domains, such as academic discourse, and healthcare. However, the tourism sector remains largely unexplored in this regard, despite evidence suggesting that AI can provide personalized and efficient travel solutions (K□rt□/Aşkun 2021). In spite of the plethora of studies in tourism discourse (e.g., Dann 1996; Cappelli 2006; Heller et al. 2014; Maci 2017), those specifically addressing sustainable and inclusive practices are relatively scarce (Malavasi 2017; Caimotto 2020; Lazzeretti 2021; Pato et al. 2021; Pasquini 2018; Cappelli/Masi 2019).

This paper aims at filling these research gaps, examining a corpus of texts belonging to various textual genres within tourism discourse (including information leaflets and brochures), created by different AI text generators (e.g., ChatGPT and Gemini). Given the foci of the research, the selected texts provide information about some of the most famous trails in Europe as defined in the ranking of international walking tour organisations, since this type of tourism is generally considered more environmentally sustainable. Quantitative and qualitative methods of corpus-assisted discourse studies (Partington et al. 2013), are adopted to shed light on whether AI-generated tourism texts both fulfill the communicative functions typical of tourism discourse and adopt a sustainable and inclusive approach. The findings offer actionable insights, including tailored prompt strategies for generating texts aligned with DEI and sustainability goals, in order to be strategically used by local institutions and enterprises.

List of references

Caimotto, M. C. (2020). Discourses of cycling, road users and sustainability: An ecolinguistic investigation. Palgrave Macmillan.

Cappelli, G. (2006). The translation of tourism discourse: Cultural representations in translation. Peter Lang.

Cappelli, G., & Masi, S. (2019). Knowledge dissemination through tourist guidebooks: Popularization strategies in English and Italian guidebooks for adults and for children. In: M. Bondi, S. Cacchiani, & S. Cavalieri (Eds.), Communicating Specialized Knowledge: Old Genres and New Media (pp. 124-161). Cambridge Scholars Publishing.

Dann, G.M.S. (1996). The Language of Tourism. A Sociolinguistic Perspective. Oxon: CAB International.

Lazzeretti, C. (2021). Communicating Sustainable Tourism in English and Italian: A Contrastive Analysis. Linguæ &-Rivista di lingue e culture moderne, 19(2), 133-154.

Heller, M., Pujolar, J., & Duchêne, A. (2014). Linguistic commodification in tourism. Journal of Sociolinguistics, 18(4), 539-566.

K□rt□, İ. G., & Aşkun, V. (2021). Artificial intelligence in tourism: A review and bibliometrics research. Advances in Hospitality and Tourism Research (AHTR), 9(1), 205-233.

Maci, S. M. (2020). English tourism discourse: insights into the professional, promotional and digital language of tourism. Hoepli.

Malavasi, D. (2017). “No one can be the invisible tourist-but we like that you are trying”: An Analysis of the Language of Sustainable Tourism. In: M. Gotti, S. Maci, & M. Sala (Eds.), Ways of Seeing, Ways of Being: Representing the Voices of Tourism (pp. 363-377). LINGUISTIC INSIGHTS, 228. Peter Lang.

Partington, A., Duguid, A., & Taylor, C. (2013). Patterns and meanings in discourse: Theory and practice in corpus-assisted discourse studies (CADS). Amsterdam: John Benjamins Publishing Company.

Pasquini, E. (2018). Politically correct tourism discourse in airport websites guidelines for inclusive travelling. Scripta Manent, 12(1), 21-37.

Pato, M. L., & Duque, A. S. (2021). Sustainability communication in rural tourism: Website content analysis, in viseu dão lafões region (Portugal). Sustainability, 13(16), 8849.

Keys to populism? Exploring varieties of mainstream populist discourse through key item analysis

Natalia Zawadzka-Paluektau, Witold Kieraś

Institute of Computer Science, Polish Academy of Sciences

Populism is simultaneously one of the most widely used and most poorly understood concepts relating to the language of politics. In fact, it is so frequently put into question that a whole new level of meta-reflexivity has emerged, where it is now customary to 'acknowledge the acknowledgment' of the contested nature of populism (Moffitt & Tormey, 2014, p. 382). This study aims to contribute to its understanding and to answer calls (Demata et al., 2020, p. 12) for the development of a more rigorous and comprehensive methodological framework of identifying populist discourse in large collections of texts by drawing on the Corpus-Assisted Discourse Analysis toolkit, specifically key item analysis. In order to test the 'populist Zeitgeist' hypothesis (Mudde, 2007), and to fill in the gap in research on populism in Eastern Europe (Zienkowski & Breeze, 2019), this methodology is applied to a custom-built corpus of Polish mainstream politicians' election campaign speeches.

The results of the analysis point to populist contagion across the mainstream of the political spectrum, provide counter-evidence to the claims of a transitory nature of populism by showing that the ruling party is as likely to employ populist rhetoric as the opposition, and reveal significant variation, despite the relatively subtle ideological differences between the politicians whose speeches were examined. This suggests that populism is not a uniform and invariable discursive phenomenon characterised by a set of fixed linguistic features (as it is often treated in the literature), and that the commonly-held distinction between left- and right-wing populism is unlikely to capture the full complexity of populist communication. Consequently, populism should not be considered and examined merely as an attachment to a 'host' (usually extreme) ideology, but rather as a complex discursive phenomenon in its own right.

List of references

1. Demata, M., Conoscenti, M., & Stavrakakis, Y. (2020). Riding the populist wave: Metaphors of populism and anti-populism in the Daily Mail and The Guardian. Iperstoria, 15, 8–35.
2. Mudde, C. (2007). Populist radical right parties in Europe. Cambridge University Press.
3. Moffitt, B., & Tormey, S. (2014). Rethinking populism: Politics, mediatisation and political style. Political Studies, 62(2), 381–397.
4. Zienkowski, J., & Breeze, R. (2019). Introduction: Imagining populism and the peoples of Europe. In J. Zienkowski & R. Breeze (Eds.), Imagining the peoples of Europe: Populist discourses across the political spectrum (pp. 1–18). John Benjamins Publishing Company.

Aboutness in English legal decisions: Choosing between key words and n-grams

Xiao Zhang

Xi'an International Studies University

This study presents a comparative analysis of the aboutness in English legal decisions through investigating key items. Aboutness is an abstract concept that deals with the topic or main content of a text and it is intrinsically linked to the keyness of key items, which include key words and n-grams. The Keyness score indicates the degree of relevance of one key item to the topic of a text, i.e. the degree of aboutness. The aims of this study are twofold: 1) to measure the keyness of n-grams and discuss the possible aboutness revealed by them; 2) to discuss suitable indicators of aboutness.

To address these aims, Corpus of Contemporary English Legal Decisions (CoCELD), comprising 288 texts of British judicial decisions from 1950 to 2021 (totaling 733,227 words), served as the target corpus. BNC, LOB and F-LOB were combined to form the reference corpus.

Key words, 3-grams and 4-grams were extracted from the corpora firstly. AntConc was used for calculating the keyness of single key words, while a Python tool was developed to extend keyness analysis to n-grams. Consistent with previous studies, log-likelihood was used to measure the keyness of single words. However, while most prior research on n-grams relies on raw frequencies, with only a few exploring simple statistics for ranking n-grams (e.g., Andersen 2016; Giampieri 2023), this study proposes applying the same statistical measures used for key words to n-grams, arguing that they are statistically comparable. Adjustments were made to the raw frequencies in this measurement.

By comparing the aboutness revealed by key items, preliminary results suggest that single key words may be more effective indicators of the aboutness in legal texts than n-grams. Overall, this study provides insights into exploring the aboutness and highlights the necessity for further discussion on determining suitable indicators of aboutness in other genres.

List of references

Andersen, G. (2016). Using the corpus-driven method to chart discourse-pragmatic change. In H. Pichler (Ed.), Discourse-Pragmatic variation and change in English: New methods and insights (pp.21-40). Cambridge: Cambridge University Press.

Anthony, L. (2024). AntConc (Version 4.3.1) [Computer Software]. Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software/AntConc

Baker, P. (2004). Querying keywords: Questions of difference, frequency, and sense in keywords analysis. Journal of English Linguistics, 32(4), 346-359.

Barlow, M. (2013). Individual differences and usage-based grammar. International Journal of Corpus Linguistics, 18(4), 443-478.

Bondi, M. & Scott, M. (Eds.). (2010). Keyness in texts. Amsterdam: John Benjamins.

Breeze, R. (2013). Lexical bundles across four legal genres. International Journal of Corpus Linguistics, 18(2), 229–253.

Egbert, J. & Lee, T. R. (2024). Prototype-by-component analysis: A corpus-based, intensional approach to ordinary meaning in statutory interpretation. Applied Corpus Linguistics, 4(1), Article 100078.

Giampieri, P. (2023). Key n-Grams in EU directives and in the UK national legislation on consumer contracts. International Journal for the Semiotics of Law, 37, 59–75.

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

Johnson, A. & Wright, D. (2014). Identifying idiolect in forensic authorship attribution: An n-gram textbite approach. Language and Law, 1(1), 37-69.

Makouar, N., Devine, L. & Parker, S. (2023). Legislating to control online hate speech: A corpus-assisted semantic analysis of French parliamentary debates. International Journal for the Semiotics of Law, 36, 2323–2353.

Pojanapunya, P. & Todd, R. W. (2016). Log-likelihood and odds ratio: Keyness statistics for different purposes of keyword analysis. Corpus Linguistics and Linguistic Theory, aop: DOI 10.1515/cllt-2015-0030.

Rayson, P. (2019) Corpus analysis of key words. In C. A. Chapelle (Ed.), The Concise Encyclopedia of Applied Linguistics (pp. 320-326). New Jersey: Wiley-Blackwell.

Rayson, P. & Potts, A. (2020). Analysing keyword lists. In M. Paquot & S. Th. Gries (Eds.), A Practical handbook of corpus linguistics (pp.119–139). Berlin & New York: Springer.
Rodríguez-Puente, P. & Hernández-Coalla, D. (2023). The Corpus of Contemporary English Legal Decisions, 1950–2021 (CoCELD): A new tool for analysing recent changes in English legal discourse. ICAME Journal, (47), 109 - 117.
Scott, M. (1997). PC analysis of key words — And key key Words. System, (25), 233-245.
Scott, M. & Tribble, C. (2006). Textual patterns: Key words and corpus analysis in language education. Amsterdam: John Benjamins.
Shinichiro, T. (2017). Multi-Word sequences in legal discourse. Language, Culture, and Communication, 9, 113-147.
Stubbs, M. (1996). Text and corpus analysis: Computer assisted studies of language and culture. Oxford: Blackwell.
Stubbs, M. (2002). Words and phrases: Corpus studies of lexical semantics. Oxford: Blackwell.
Stubbs, M. (2010). Three concepts of keywords. In M. Bondi & M. Scott (Eds.), Keyness in texts (pp. 21–42). Amsterdam: John Benjamins.
Taylor, C. (2013). Searching for similarity using corpus-assisted discourse studies. Corpora, 8(1), 81–113.
Wright, D. (2017). Using word n-grams to identify authors and idiolects: A corpus approach to a forensic linguistic problem. International Journal of Corpus Linguistics, 22(2), 212–241.
Wright, D. (2021) Corpus approaches to forensic linguistics. In M. Coulthard, A. May & R. Sousa-Silva (Eds.), The Routledge Handbook of Forensic Linguistics (2nd ed., pp. 611-627). London: Routledge.
Wulff, S. & Gries, S. Th. (2024). CLLT 'versus' Corpora and IJCL: a (half serious) keyness analysis. Corpus Linguistics and Lingustic Theory, aop: https://doi.org/10.1515/cllt-2024-0050.
Corpora
Rodríguez-Puente, P. & Hernández-Coalla, D. (2022). Corpus of Contemporary English Legal Decisions, 1950-2021 (CoCELD). Oviedo: University of Oviedo. https://varieng.helsinki.fi/CoRD/corpora/CoCELD/
British National Corpus (BNC). http://www.natcorp.ox.ac.uk/
The Freiburg–LOB Corpus of British English (F-LOB). https://varieng.helsinki.fi/CoRD/corpora/FLOB/
The Lancaster-Oslo/Bergen Corpus (LOB). https://varieng.helsinki.fi/CoRD/corpora/LOB

A corpus-based critical discourse analysis on the discursive construction of national identity through representation of political figures in the COVID-19 discourse in Chinese newspapers

Xiaowen Zhao

University of Birmingham

Background and Research question

National identity refers to national members' sense of belonging to a particular national group. It is dynamic and can often be aroused by crisis in troubled times (Wodak et al. 2009) such as the global pandemic COVID-19. In this study, I adopted corpus-based critical discourse analysis (CDA) to analyze COVID-19 news articles on the Mandarin-language Xinhua Daily Telegraph (XDT) and the English-language China Daily (CD) from 27 December 2019 to 28 April 2020, and explored how Chinese national identities are constructed through the portrayal of political social actors in the two corpora.

Data and methods

The analysis is done by examining the collocates and concordances of political social actors in the XDT corpus and CD corpus generated from keyword analysis. The concordance analysis is carried out to examine the patterns of use of these collocates in co-text. The analytical frameworks used for CDA are the three-tier framework developed specifically for the analysis on national identity (Wodak et al., 2009) and the 'representation of social actors' model.

Results

XDT demonstrates a tendency to emphasize political figures associated with the Chinese Communist party, where offensive war metaphors are frequently employed to portray party members as leading warriors in the fight against the pandemic. A hierarchical 'government – leading cadres – primary-level cadres – general public' power structure is constructed, enabling the efficient distribution of authority to facilitate epidemic prevention and control. Conversely, in CD, greater emphasis is placed on promoting international cooperation, with China positioned as a global leader contributing to the containment of the pandemic.

Implications

This study helps readers to understand how Chinese state media can potentially influence the public perception and acceptance of government mandates through the portrayal of the party, government, and ordinary people's role in handling the pandemic, to advocate regime legitimacy and promote national solidarity.

List of references

Wodak Ruth, Rudolf De Cillia, Reisigl Martin, Liebhart Karin, Hirsch Aron, and Rodger Ruth. 2009. The Discursive Construction of National Identity (2nd ed.). Edinburgh University Press.

Achieving communicative purposes: A cross-linguistic and corpus-based analysis of British and Chinese conversations

Lily Zhou, Tony McEnery

Lancaster University

Conversational discourse is often examined at a micro-level, focusing on units such as turns and speech acts. However, at a macro level, it comprises larger units. This study is based on the framework developed by Egbert et al. (2021), which segments English conversational discourse into discourse units and classifies them into nine communicative purposes. This research applies this framework to three corpora, all containing conversations structured according to the GESE exam (Graded Examinations in Spoken English). The first corpus is a Mandarin Chinese conversation corpus, compiled specifically for this study through conversations with 100 native Mandarin speakers. The second corpus is a British conversation corpus compiled by Fox (2024), consisting of conversations between native British speakers. The third corpus is a subset of the Trinity Lancaster Corpus (Gablasova et al., 2019), which includes conversations between British examiners and Chinese learners of English participating in the GESE exam. With the three corpora segmented and annotated using Egbert et al.'s (2021) framework, this research utilises keyword analysis to explore lexical items typical for fulfilling each communicative purpose across the three corpora. These lexical items will be examined in context to understand their syntactic, semantic, or pragmatic usage at the micro level, and how they achieve overarching communicative purposes at the macro level. Based on the identified lexical items and the statistics for each communicative purpose, this paper will compare and contrast the differences and similarities among conversations in Mandarin Chinese, British English, and those of Chinese English learners. Ultimately, this research will illustrate whether first language interference occurs in the conversations of Chinese English learners, impacting the realisation of discourse units and communicative purposes. The findings will have implications for understanding how communicative purposes differ between these two languages and for English language teaching in effectively achieving communicative purposes in conversations.

List of references

Egbert, J., Wizner, S., Keller, D., Biber, D., McEnery, T., & Baker, P. (2021). Identifying and describing functional discourse units in the BNC Spoken 2014. Text & Talk, 41(5-6), 715-737.

Fox, L. (2024). Verb+ noun collocations in L1 and L2 English spoken language examinations: Introducing the Trinity Lancaster Corpus of L1 Spoken English to investigate formulaic language. (Doctoral dissertation, Lancaster University). CORE. https://core.ac.uk/download/pdf/603233295.pdf

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

A study on the production of Korean relative clauses by L2 learners: An analysis based on learner corpora

Sixuan Zhu, Xuehui Cao

Seoul National University

Previous studies on Korean learners' acquisition of relative clauses (RCs) have predominantly examined syntactic factors, consistently indicating that subject relative clauses (SRCs) are easier to process than object relative clauses (ORCs) (O'Grady et al., 2003; Huh, 2015; Ju, 2024), aligning with the Noun Phrase Accessibility Hierarchy (NPAH; Keenan & Comrie, 1977). Emerging evidence, however, points to potential interactions between syntactic configurations and semantic properties, particularly head noun animacy. Nevertheless, investigations into animacy effects in Korean RCs remain notably limited, with existing work predominantly concentrating on verbal predicates.

This study addresses this gap by analyzing 288 essays by Chinese, English, and Japanese learners from the National Institute of the Korean Language Learner Corpus. Using Park's (2019) definition of RCs, each instance was systematically annotated for RC type, head noun animacy, and predicate category.

Results show that learners most frequently produced subject RCs, followed by direct object and oblique RCs, supporting the NPAH, although only subject RCs were significantly more frequent than other types.

Inanimate head nouns appeared significantly more often than animate ones. A detailed analysis indicated that animacy interacted with syntactic role only when the predicate was a transitive verb—where subject-object distinctions are structurally relevant. Notably, learners favored inanimate direct object RCs (DO+I) and animate subject RCs (SU+A), with DO+I occurring more frequently. This pattern held across all three L1 groups and was significant only in transitive verb (VT) contexts.

L1 background also influenced RC usage. Japanese learners produced the most RCs and used the widest range of syntactic structures and predicate types, though overall counts did not significantly differ from Chinese and English learners. Semantically, Chinese learners most frequently used inanimate head nouns, significantly more than Japanese learners. Predicate type further affected animacy: learners preferred inanimate head nouns with adjectival predicates, decreasing across copular, intransitive, and transitive predicates.

List of references

Huh, S. (2015). A corpus study of L2 Korean relative clause development. Language Research (어학연구).

Keenan, E. L., & Comrie, B. (1977). Noun phrase accessibility and universal grammar. Linguistic Inquiry, 8(1), 63–99.

O'Grady, W., Lee, M., & Choo, M. (2003). A subject–object asymmetry in the acquisition of relative clauses in Korean as a second language. Studies in Second Language Acquisition, 25(3), 433–448.

Park, H. J. (박형진). (2019). The syntax and semantics of internally-headed relative clauses in Korean. Bangyo Korean Language and Literature Studies (반교어문연구), 52, 87–118.

Zhu, S. (2024). A study of Chinese-speaking learners' processing of Korean relative clauses and use of linguistic cues: Based on a self-paced reading task (Master's thesis). Seoul National University.

PRE-RECORDED PRESENTATIONS

Bridging cultures on screen: Corpus linguistic insights into intercultural discourse in Saudi Arabian filmmaking

Banan Assiri

King Khalid university

In an increasingly globalised world, intercultural communication skills are essential—particularly in creative industries like filmmaking, where narrative construction and cultural representation are central to audience engagement (Belfiore, 2020). This empirical study investigates the integration of intercultural communication within the institutional discourse surrounding filmmaking in Saudi Arabia, employing a corpus-assisted approach to examine contemporary educational materials and industry-produced content. Through a comprehensive corpus-assisted discourse analysis (Gillings, Mautner and Baker, 2023), this research critically examines how intercultural themes are represented, negotiated, and operationalised, illuminating both systemic gaps and emergent strengths in institutional practices (Hofstede, 2001; Handford and Koester, 2024).

By conducting a detailed analysis of linguistic patterns, discursive genres, and cultural markers, this investigation reveals how the Saudi Arabian filmmaking industry conceptualises and implements strategies to engage with diverse global audiences and narratives. Initial findings indicate significant variations in intercultural awareness across institutional contexts, with educational materials emphasising traditional cultural values while industry documents demonstrate a more hybridised approach to cultural representation. The corpus analysis reveals distinct linguistic patterns that suggest an emerging 'third space' in Saudi filmmaking discourse, where traditional narrative forms are being reimagined through a contemporary global lens. The findings make a significant contribution to the broader discourse on industrial communication, intercultural competence development, and the methodological applications of corpus linguistics in examining cultural exchange mechanisms (Hua, 2018). Specifically, the study offers novel insights into the distinctive socio-cultural landscape of Saudi Arabia's emerging film sector and its influence on professional development frameworks, thereby extending current theoretical models of intercultural competence (Spencer-Oatey and Franklin, 2009), and addressing the crucial intersection between theoretical foundations and industry practices in the contemporary intercultural creative industry.

List of references

Belfiore, E., 2020. Whose cultural value? Representation, power and creative industries. International journal of cultural policy, 26(3), pp.383-397.

Byram, M. and Masuhara, H., 2013. Intercultural competence. Applied linguistics and materials development, pp.143-159.

Gillings, M., Mautner, G. and Baker, P., 2023. Corpus-assisted discourse studies. Cambridge University Press.

Handford, M. and Koester, A., 2024. Language and creativity at work: A corpus-assisted model of creative workplace discourse. Taylor & Francis.

Hofstede, G., 2001. Culture's consequences: Comparing values, behaviors, institutions and organizations across nations. Thousand Oaks.

Hua, Z., 2018. Exploring intercultural communication: Language in action. Routledge.

Spencer-Oatey, H., & Franklin, P., 2009. Intercultural Interaction: A Multidisciplinary Approach to Intercultural Communication. Basingstoke: Palgrave Macmillan.

Business communication in recruitment: A CADS approach to job advertisements

Francesca Nannetti, Matteo Di Cristofaro

University of Modena and Reggio Emilia

In light of the vast *war for talent* (Osorio & Madero, in press), effective recruitment represents a strategic tool for attracting desired candidates (De Cooman & Pepermans, 2012). Most often, job advertisements are the initial point of contact between employers and job seekers, especially young graduates entering the labour market for the first time (Breeze, 2013). As one of the first sources of information about a company, job advertisements not only notify about vacancies, but also express how an organisation wants to be perceived by potential applicants by conveying its values, culture and identity (Rafaeli & Oliver). Assuming that organisations are constituted and sustained by texts and discursive practices couched in a variety of genres, which do not merely accompany organisational life but actually constitutes it (Mautner, 2021), job advertisements can be analysed as a form of business communication, with characteristic discursive features (Breeze, 2013). This research employs a CADS approach (Gillings et al., 2023), which links "micro-linguistic choices with context, not only on a text level, but also on organizational, political, and societal levels" (Gillings et al., 2024: 4) to examine a corpus of job advertisements targeting Generation Z graduates. To explore how companies encourage candidates to believe the claims made in the advertisement and to apply for the position (Pourfarhad, 2012), a purpose-built specialised corpus (Jaworska, 2017) of online job advertisements has been collected from a job offers notice board accessible to recent graduates of Italian universities. Job advertisements (1859) were collected from October 2023 to October 2024. Extraction and formatting of the data was conducted using a custom Python script, preserving both metadata and textual content. The corpus was loaded and analysed on #LancsBox X (Brezina & Platt, 2024). Results provide insights into how companies approach job advertising for Generation Z and communicate organisational culture and identity.

List of references

Breeze, R. (2013). Corporate discourse. Bloomsbury.

Brezina, V. & Platt, W. (2024). #LancsBox X [software]. Lancaster University. URL: http://lancsbox.lancs.ac.uk.

De Cooman, R., & Pepermans, R. (2012). Portraying fitting values in job advertisements. Personnel review, 41(2), 216-232.

Gillings, M., Mautner, G., & Baker, P. (2023). Corpus-assisted discourse studies. Cambridge University Press.

Gillings, M., Learmonth M., & Mautner, G. (2024). Taking the Road Less Travelled: How Corpus-Assisted Discourse Studies Can Enrich Qualitative Explorations of Large Textual Datasets. British Journal of Management, 35, 1-13.

Jaworska, S. (2017). Corpora and corpus linguistic approaches to studying business language. In G. Mautner & R. Franz (Eds.), Handbook of business communication: Linguistic approaches, 426–443. Walter de Gruyter.

Mautner, G. (2020). Business discourse. In Friginal, E., & Hardy, J. A. (Eds.), The Routledge Handbook of Corpus Approaches to Discourse Analysis, 319-333. Routledge.

Osorio, M.L., & Madero, S. (in press). Explaining Gen Z's desire for hybrid work in corporate, family, and entrepreneurial settings. Business Horizons.

Pourfarhad, M. (2012). Introducing Communicative Event As A Tool To Communicate Via the Medium Of Language: The Case of Job Advertisement. International Journal of Applied Linguistics & English Literature 1(1), 104-112.

Rafaeli, A., & Oliver, A. L. (1998). Employment ads: A configurational research agenda. Journal of management inquiry, 7(4), 342-358.

A corpus-driven comparative investigation of academic discourse produced by AI versus humans

Juan Shao

Xi'an Jiaotong University

With the growing integration of artificial intelligence (AI) into the realm of education, its influence on academic writing has attracted considerable attention. There is a widely held belief that AI can effectively replace humans in language-related tasks and past research on Generative AI has mainly focused on evaluating the quality of content it produces. However, there has been a lack of in-depth investigation into the linguistic patterns in academic discourse produced by AI versus human authors. The aim of this study was to assess the degree of similarity between academic texts produced by artificial intelligence (GPT) and those written by humans. We constructed two specialized corpora consisting of academic texts in the fields of linguistics and medicine, generated by ChatGPT and selected the Hu-LM Corpus for comparison. Sketch Engine was utilized to facilitate the linguistic analysis of the two corpora, including part-of-speech analysis, N-grams (formulaic sequences), and concordance analysis. The results showed significant differences between AI-generated and human-authored texts. First, ChatGPT's output contained some non-academic lexical items and made extensive use of figurative language, indicating a limitation in the model's ability to select contextually appropriate vocabulary. Moreover, while ChatGPT's output featured common formulaic sequences found in academic English, the concordance analysis revealed its tendency to use syntactically and semantically equivalent structures generated through synonym substitution. These behaviors can be attributed to the architectural limitations and probabilistic mechanisms underlying large language models. This study contributes to the linguistic analysis of AI-generated academic texts and enriches corpus linguistics research by incorporating the output of large language models into the analytical framework. It also lays a foundation for further refining the model's ability to generate contextually appropriate language output.

List of references

Ang, L.H., Tan, K.H., 2018. Specificity in English for academic purposes (EAP): A corpus analysis of lexical bundles in academic writing. 3L: Lang. Linguist. Lit. 24 (2), 82–94.

Baidoo-Anu, D., Owusu Ansah, L., 2023. Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. J. AI 7 (1), 52–62.

Berber Sardinha, T., 2024. AI-generated vs human-authored texts: A multidimensional comparison. Appl. Corpus Linguist. 4, (1) 100083.

Pérez-Llantada, C., 2014. Formulaic language in L1 and L2 expert academic writing: Convergent and divergent usage. J. Engl. Acad. Purp. 14, 84–94.

Tudino, G., Qin, Y. 2024. A corpus-driven comparative analysis of AI in academic discourse: Investigating ChatGPT-generated academic texts in social sciences. Lingua 312, 103838.

Zindela, N., 2023. Comparing measures of syntactic and lexical complexity in artificial intelligence and L2 human-generated argumentative essays. Int. J. Educ. Dev. Using Inf. Commun. Technol. 19 (3), 50–68.

POSTER PRESENTATIONS

Bridging the Data-Driven Learning Gap for young ESL learners

Cansu Akan

Chemnitz University of Technology

This ongoing research project addresses the significant gap in Data-Driven Learning (DDL) studies focusing on young ESL learners. DDL, as defined by Johns (2002), aims to confront "the learner as directly as possible with the data, and to make the learner a linguistic researcher" (p. 108). While DDL has shown promise in language education, it has primarily been utilized with adult or teenage learners due to the complexity of existing corpus tools. Therefore, this project responds to the observation that "data driven learning with young learners is still very much underexplored" (Crosthwaite & Baisa, 2023).

The project involves the development of a multi-modal corpus application for ESL learners aged 8–10. The study is carried out in Germany with 30 primary school students with varying L1s. The corpus is compiled from a selection of CEFR A1-A2+ graded readers which are age and theme appropriate for the target group. The learning application employs serious game design principles and consists of DDL vocabulary activities catered for young learners. The multi-modal perspective is added via the integration of visuals, audio, characters and interactive haptic-tools.

This study employs a pre- and post-tests design along with semi-structured interviews with learners and teachers. The data from the learning app includes activity logs such as usage frequency and time on task. The data is triangulaised by cross-referencing test results, interaction logs, and interview responses to provide a comprehensive analysis of how young learners engage with gamified DDL activities and whether DDL strategies can effectively be integrated into young learner ESL curriculum through the use of online multi-modal corpus teaching resources.

The findings will contribute to corpus linguistics, second language acquisition, computer-assisted-language-learning and curriculum design research and provide a foundation for broader educational use, potentially leading to "long-lasting DDL use" in young learner settings (Crosthwaite & Baisa, 2023).

List of references

Crosthwaite, P. & Baisa, V. (2023). Generative AI and the end of corpus-assisted data- driven learning? Not so fast!. Applied Corpus Linguistics. 3.

Johns, Tim F. (2002). "Data-driven learning: The perpetual challenge." Bernhard Kettemann and Georg Marko, eds. Teaching and Learning by Doing Corpus Analysis. Proceedings of the Fourth International Conference on Teaching and Language Corpora, Graz 19-24 July, 2000. Amsterdam: Rodopi, 107-117.

Developing a computational lexicon for the multifaceted concept of "quality of life" In Quranic discourse- A Quranic Arabic corpus-based approach

Jawharah Saeed Alasmari

Princess Nurah university

This research aims to develop a lexicon for the concept of "quality of life" in the Quran by examining a corpus of 77,439 Arabic words. The study seeks to identify key semantic indicators associated with quality of life, such as "justice, progress, mercy, and benevolence," as a framework for constructing a comprehensive understanding of quality of life from a Quranic perspective. A total of 119 primary semantic words were identified. The research explores questions related to the Quranic terms representing quality of life, the interplay of semantic indicators in shaping this concept, and the effectiveness of text mining and machine learning techniques in classifying Quranic discourse. The methodology involves extracting key semantic words using specialized corpus analysis computational tools, analyzing them within their textual contexts to determine their semantic associations, and categorizing them based on different recipient groups (women, men, children, society). The study also benefits from constructing a semantic map to illustrate the relationships between concepts and values related to quality of life and applying machine learning algorithms to analyze semantic patterns in Quranic texts. The research analyzes word frequency and classifies sentences based on social, environmental, and psychological conditions, contributing to a deeper understanding of this concept in Quranic discourse. Preliminary results indicate that environmental laws were the most frequent, followed by social relations, while laws of worship were the least frequent. Confusion matrix analysis revealed high accuracy in classifications, with some overlap between semantic indicators. The research is expected to result in the development of a comprehensive conceptual map illustrating the relationships between values and concepts related to quality of life, and the construction of a classification model that enables the application of the results to improve quality of life according to the Quranic perspective. Furthermore, the study will introduce new methodological tools for analyzing religious texts.

List of references

Levin, B. “Semantics and pragmatics of argument alternations.” Annu. Rev. Linguist., vol.1, no.1, 2015, pp. 63-83. https://doi.org/10.1146/annurev-linguist-030514-125141 (2) Leech, Geoffrey. Semantics The Study of Meaning. 1981. Second ed. Great Britain: Penguin Books. (3) Arad, M. “A minimalist view of the syntax-lexical semantics interface.” University College of London Working Papers in Linguistics, 1996, vol. 8, pp. 215-242. https:// www.phon.ucl.ac.uk/home/PUB/WPL/96papers/arad.pdf 42/169 A Quranic Arabic Corpus-based Quality-of-Life Pragmatic Analysis 26 (4) Stringer, D. “Lexical semantics: Relativity and transfer.” IGI Global, In Applied lin- guistics for teachers of culturally and linguistically diverse learners.2019, pp. 180-
https://doi.org/10.4018/978-1-5225-8467-4.ch007 (5) Kelly, B. T., & Bhangal, N. K. “Life narratives as a pedagogy for cultivating critical self-reflection.” New Directions for Student Leadership, vol. 2018, no. 159, 2018, pp. 41-52. https://doi.org/10.1002/yd.20296 (6) Geeraerts, D, 'Lexicography and Theories of Lexical Semantics', in Philip Durkin (ed.), The Oxford Handbook of Lexicography. 2015. (online edn, Oxford Academic, 7 Mar. 2016), https://doi.org/10.1093/oxfordhb/9780199691630.013.32, accessed 28 Sept. 2023. (7) Alsina, A. (2007). Beth Levin & Malka Rappaport Hovav, Argument realization (Research Surveys in Linguistics). Journal of Linguistics - J LINGUIST. 43. 10.1017/S0022226707004677.

Hua-Mei, Chen., Erik, P., Blasch., Nichole, Sullivan., Genshe, Chen. (2022). Trajectory-Based Pattern of Life Analysis. 2591-2595. doi: 10.1109/ICIP46576.2022.9897585

Joseph, A., Mayo. (2001). Life analysis: using life-story narratives in teaching life-span developmental psychology. Journal of Constructivist Psychology, doi: 10.1080/10720530125850

Karen, C., Quackenbush., Don, A., Quackenbush. (2019). Stylistics Analysis of Psalm of Life. International journal of scientific and research publications, 9(3):8723-. doi: 10.29322/IJSRP.9.03. 2019.P8723

Maya, Arad. (1996). A minimalist view of the syntax-lexical semantics interface.

Pollice, B., Thiel, C., Baratz, Me. (2022). Life Cycle Analysis. Operative Techniques in Orthopaedics, doi: 10.1016/j.oto.2022.100998

Tiri, Bergesen, Schei. (2013). Everyday Life Discourses in Kindergarten. Cultural-Historical Psychology, 9(2):31-37.

URL: https://doi.org/10.1146/annurev-linguist-030514-125141

From Girls to #Thatgirls: Popular feminism and the representation of femininity on TikTok

Selenia Anastasi

University of Rome La Sapienza, University of Genoa

Since 2018, TikTok firmly established itself as a key platform in contemporary youth digital culture, culminating in its status as the most visited web domain worldwide in 2022. Labelled a “hub for girls” (Banet-Weiser & Maddocks, 2023), TikTok effectively translates the “girls’ bedroom culture” (McRobbie & Garber, 1976) into a dynamic online ecosystem. However, corpus-based studies focusing on TikTok remain scarce (Raffone, 2022; Donati, 2023), and the circulation of internalised sexism on social media is still empirically underexplored (Rosida et al., 2022; Patouras & Tanner, 2024).This study addresses these gaps by examining a corpus of 50 Italian TikTok videos (21,656 tokens) tagged with the trending hashtag #thatgirl, selected according to popularity indicators (views, likes, comments). Creators' identities were anonymised, and the textual corpus was compiled using the GPT 4-omni speech-to-text model to entirely transcribe the spoken content. Multimodal features are annotated considering three levels: background sounds, salient visual elements such as setting, and framing of the subject. By integrating corpus-based and thematic approaches, I first identified keywords and their principal collocations (Baker, 2006) using SketchEngine, then conducted an iterative thematic analysis (Braun & Clarke, 2006) through concordance lines to generate codes reflecting dominant themes. The final thematic categories were established by cross-referencing these codes with quantitative findings and interpreted within a feminist discourse analysis framework (Lazar, 2007, 2011; Mills & Mullany, 2011; Formato 2024). In alignment with TikTok’s algorithmic logic, which predominantly rewards conventional beauty standards, a significant proportion of the content depicts female subjects engaged in stereotypically feminine activities, such as beauty routine and performing domestic tasks. Such content revolves around the romanticisation of domestic work, popular feminist notions of empowerment (Banet-Weiser 2018), and men-women relationship advice, thereby underscoring the need to critically examine how internalised sexism operates (and is reinforced) through mainstream social media platforms.

List of references

Banet-Weiser, S. (2018). Empowered: Popular Feminism and Popular Misogyny. Duke UP.
Banet-Weiser, S., & Maddocks, S. (2023). Networked misogyny on TikTok: A critical conjuncture. In The Routledge Companion to Gender, Media and Violence (pp. 369-379). Routledge.
Baker, P. (2006) Using Corpora in Discourse Analysis. London: Continuum
Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative research in psychology, 3(2), 77-101.
Donati, M. (2023). Analysing Eating Disorder Discourse in the Italian TikTok Community: a Corpus-based Approach, Master Dissertation, University of Bologna.
Formato, F. (2024). Feminism, corpus-assisted research and language inclusivity. Cambridge University Press.
Lazar, M. M. (2007). Feminist critical discourse analysis: Articulating a feminist discourse praxis. Critical discourse studies, 4(2), 141-164.
Lazar, M. M. (2011). The right to be beautiful: Postfeminist identity and consumer beauty advertising. In New femininities: Postfeminism, neoliberalism and subjectivity (pp. 37-51). London: Palgrave Macmillan UK.
McRobbie, A., & Garber, J. (1976). Girls and subcultures. In S. Hall & T. Jefferson (Eds.), Resistance through rituals: Youth subcultures in post-war Britain (pp- 209-222). London: Harper Collins Academics.
Mills, S., & Mullany, L. (2011). Language, gender and feminism: Theory, methodology and practice. Routledge.
Patouras, S., & Tanner, C. (2024). “Oh how I love being a woman”: post-feminism and contemporary femininity on TikTok. Feminist Media Studies, 1-18.
Raffone, A. (2022). “Her leg didn’t fully load in”: A digitally-mediated social-semiotic critical discourse analysis of disability hate speech on TikTok. International journal of language studies, 16(4).
Rosida, I., Ghazali, M. M., Dedi, D., & Salsabila, F. S. (2022). The manifestation of internalized sexism in the pick me girl trend on TikTok. Alphabet: A Biannual Academic Journal on Language, Literary, and Cultural Studies, 5(1), 8-19.

Building the Corpus of American Style Guides (CASG): Student-led innovative methods for collection, correction, and coding

Holly Baker

Brigham Young University

Linguistic prescriptivism is an emerging discipline in the field of linguistics (Beal, Luckac, & Straaijer, 2023) and often involves corpus-based approaches to research (Szmrecsanyi & Bloemen, 2023). As a prescriptive tool, style guides, which have increasingly included prescriptions on grammar, usage, and stylistic prose, have served as a significant influence on editorial decision-making since the early 20th century and continue to impact editorial practices. In studying the work of editors and the role of style guides, however, we should keep in mind what Linda Pilliere (2020) said about that relationship: "How far copy editors reflect the values of style and usage guides in their decisions … is difficult to measure." Nevertheless, having a database of style guides may enable researchers to ask and answer questions they were not able to before in the absence of such a resource.

Currently, no corpus of style guides exists. Some researchers (see, for example, Chang & Swales, 1999; Bennett, 2009) have compiled their own limited corpora, though a more comprehensive, up-to-date database will be necessary in facilitating data collection and analysis of the role of style guides in language studies broadly. In this poster presentation, I will report on the early work of creating the Corpus of American Style Guides (CASG). Specifically, I will present data on how graduate and undergraduate research assistants (RAs), through intentional mentorship, become key players and decision-makers in the corpus construction process. Results will include weekly mentoring session topics and student reports, RA-led problem-solving efforts, peer teaching and accountability, and the creation of an RA-generated procedures manual. Ultimately, I aim to show how professors and students can work together most effectively to construct new corpora and design innovative methods for collection, correction, and coding.

Influencing the Court: Amicus briefs, abortion rights, and conceptual metaphor analysis

Andrea Banicki

Carleton University

This poster presentation shares corpus construction and some preliminary findings from an early-stage corpus-assisted critical discourse study of conceptual metaphors used in the *amicus curiae* briefs filed with the United States Supreme Court regarding *Dobbs v. Jackson Women's Health Organization* (2022), the case rescinding the constitutional right to access abortion. Amicus briefs are legal documents submitted by non-litigants who are not directly involved in a court case but have a vested interest in its outcome, therefore containing rhetorical emements indicating ideological preferences (Abrams & Potts, 2024). Empirical analysis reveals that amicus participation enhances the likelihood of litigation success (Collins, 2004). One purpose of amicus briefs is to influence policy by promoting their desired outcome (Perkins, 2018). For this reason, they can be considered tools of persuasion, making metaphor a logical figure of speech to use as it is commonly associated with persuasive communication (Charteris-Black, 2004).

Research Question:

- To what extent are ideological preferences regarding the permissibility of abortion discursively constructed using conceptual metaphors in the amicus briefs filed in *Dobbs v. Jackson* (2022)?

This corpus consists of all the amicus curiae briefs filed with the court in this case to form the Dobbs v. Jackson Amicus Briefs (DJAB) corpus. The DJAB corpus contains 137 files totaling 1,192,089 tokens. These files are divided into two sub-corpora: one of briefs aimed at restricting abortion access and the other of briefs supporting abortion access. This mixed-methods study employs Critical Metaphor Analysis (Charteris-Black, 2004), which combines qualitative close readings of briefs and concordance lines with quantitative frequency, keyword, and collocation analysis. This investigation contributes to the growing body of research in the burgeoning field of Law and Corpus Linguistics (Egbert & Römer-Barron, 2024) to enhance knowledge of metaphor as a rhetorical strategy in legal texts aiming to influence the opinions of judges.

List of references

Abrams, J.R. & Potts, A. (2024). The rhetoric of abortion in amicus briefs. Missouri Law review, 89(2), 399-476. doi: 10.2139/ssrn.4852022

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

Collins, P. M. (2004). Friends of the court: Examining the influence of amicus curiae participation in U.S. Supreme Court litigation. Law & Society Review, 38(4), 807–832. https://doi.org/10.1111/j.0023-9216.2004.00067.x

Dobbs v. Jackson Women's Health Organization, 142 S.Ct. 2228 (2022).

Egbert, J., & Römer-Barron, U. (2024). Applying corpus linguistics to the law. Applied Corpus Linguistics, 4(2), 100093. https://doi.org/10.1016/j.acorp.2024.100093

Perkins, J. (2018). Why file? Organized interests and amicus briefs in state courts of last resort. The Justice System Journal, 39(1), 39–53.

Integrating Critical Discourse Analysis (CDA) and corpus-based methods: Advancing interdisciplinary approaches in qualitative Social Sciences research

Desirée Failla

University of Padua

The integration of Critical Discourse Analysis (CDA) and Corpus Linguistics (CL), used under the umbrella of Corpus-Assisted Discourse Studies (CADS), has increasingly evolved into an interdisciplinary practice, drawing on concepts and techniques from various academic fields. While there have been various attempts to apply CADS methodologies to research in Social Sciences, these efforts remain at an early stage in many domains. As Hodge (2012) puts it: "CDA's value as a heuristic device is underestimated and under-used". Such limited integration constrains the potential of corpus-based approaches to bridge gaps, complement findings, and address questions raised by different branches of Social Sciences. This article explores the epistemological benefits and challenges of CADS, focusing on its capacity to advance interdisciplinary research in Social Sciences. This integration not only offers new tools for examining large-scale textual datasets, thus enhancing result reliability, but it also presents CADS as an epistemological tool for a better understanding of societal issues. Specific examples of successful methodological cooperation are drawn from fields such as organizational sociology, ethnography, discursive and social psychology as well as memory and conflict studies. For example, CADS can show how discourse legitimizes policies or constructs historical consensus. Moreover, it can support emerging approaches, such as integrative complexity in discourse analysis as a predictive tool for conflict outbreak (Conway, Suedfeld and Tetlock, 2001). The main purpose of this study is to encourage scholars from both domains to engage in what has already been proven to be a promising and productive cooperation. Nonetheless, the article also highlights how this interdisciplinary approach can inform the development of novel critical corpus-based frameworks for discourse analysis, fostering more nuanced understandings of societal challenges.

List of references

Althusser, Louis. 2001 [1971]. Ideology and Ideological State Apparatuses by Louis Althusser 1969-70. New York: NYU Press.

Ainsworth, Susan and Cynthia Hardy. 2004. Critical discourse analysis and identity: why bother? Critical Discourse Studies, 1(2), 225-259.

Austin, John Langshaw. 1975 [1962]. How to Do Things with Words. New York: Clarendon Press.

Baker, Paul, Costas Gabrielatos, Majid Khosravi Nik, Michał Krzyżanowski, Tony McEnery, and Ruth Wodak. 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 & Society, 19(3). 273–306.

Baker, Paul. 2023 [2006]. Using corpora in discourse analysis, 2nd edn. London: Bloomsbury.

Bakhtin, Mikhail Mikhaĭlovich. 1981. The Dialogic Imagination: Four Essays. Austin: University of Texas Press.

Békés, Vera and Peter Suedfeld. 2019. Integrative complexity. In Virgil Zeigler-Hill and Todd K. Shackelford, Encyclopaedia of personality and individual
differences. Cham: Springer International Publishing.

Bell, Allan. 1991. The language of news media. Oxford & Cambridge: Blackwell.

Bell, Allan & Garrett Peter (eds.). 1998. Approaches to media discourse. Oxford & Malden: Blackwell.

Blommaert, Jan, and Chris Bulcaen. Critical Discourse Analysis. Annual Review of Anthropology, 29(1). 447–466.

Catalano, Theresa and Waugh R. Linda. 2020. Critical discourse analysis, critical discourse studies and beyond. Cham: Springer.

Chouliaraki, Lilie, & Norman Fairclough. 2021. Discourse in Late Modernity: Rethinking Critical Discourse Analysis. Edinburgh: Edinburgh University Press.

Conway, Lucian G., Peter Suedfeld and Philip E. Tetlock. 2007. In Christie, D. J.,Wagner, R. V., & Winter, D. A. (Eds.). Peace, Conflict, and Violence: Peace Psychology for the 21st Century. Englewood Cliffs, New Jersey: Prentice-Hall.

Del Sarto, Raffaella. 2007. Il Confine Del Consenso: La Guerra Dei Sei Giorni e La Frammentazione Della Società e Della Politica Israeliana. In Arturo Marzano and Marcella Simoni (eds.). Quaranta anni dopo: confini, barriere e limiti in Israele e Palestina, 1967-2007, 33–48. Bologna: Il Ponte.

Fairclough, Norman. 1992. Discourse and Text: Linguistic and Intertextual Analysis within Discourse Analysis. Discourse & Society, 3(2). 193–217.
Fairclough, Norman. 1995a. Critical discourse analysis: the critical study of language. New York: Longman.
Fairclough, Norman. 1995b. Media discourse. London: Edward Arnold.
Fairclough, N., & Wodak, R. (1997). Critical discourse analysis. In T. A. van Dijk (Ed.),Discourse as social interaction: Discourse studies: A multidisciplinary introduction,Vol. 2, pp. 258–284). Sage Publications, Inc.
Fairclough, Norman. 2003. Analysing Discourse: Textual Analysis for Social Research. 1st ed. London: Routledge.
Foucault, Michael. 1981. History of sexuality, Vol I. Harmondsworth: Penguin.
Foucault, Michael. 1990 [1972]. The Archaeology of knowledge and the discourse on language. New York: Random House.
Foucault, Michael. 2004 [1971]. L'ordine del discorso. E altri interventi. Torino: Einaudi.
Fowler, Roger, Robert Hodge, Gunther Kress, and Tony Trew. 1979. Language and Control. London: Routledge.
Fowler, Roger. 1991. Language in the News: Discourse and Ideology in the Press. London: Routledge.
Galtung, Johan. 2000. The task of Peace Journalism. Ethical perspectives, 7(2). 162-167.
Garzone, Giuliana, and Francesca Santulli. 2004. What Can Corpus Linguistics Do for Critical Discourse Analysis?. In Alan Partington, John Morley and Louann Haarman, Corpora and Discourse. Peter Lang.
Goffman, Erving. 1959. The presentation of self in everyday life. New York: Doubleday Anchor.
Goffman, Erving. 1974. Frame analysis: an essay on the organization of experience. Cambridge: Harvard University Press.
Gramsci, Antonio. 1983. Quaderni dal carcere. Torino: Einaudi.
Habermas, Jürgen. 1971. Knowledge and Human Interests. Boston: Beacon Press.
Halliday, Michael Alexander Kirkwood. 1978. Language as social semiotic: the social interpretation of language and meaning. Baltimore: University Park Press.
Hodge, Robert and Gunther Kress. 1993. Language as Ideology. London: Routledge.
Jenner, Bryan and Titscher Stefan. 2000. Methods of text and discourse analysis. London: Sage.
Labov, William. 1972. Language in the Inner City. Philadelphia: University of Pennsylvania Press.
La Mendola, Salvatore. 2007. Comunicare interagendo. Turin: Utet Università.
Lerner, Adam B. 2022. From the ashes of history. Collective trauma and the making of International Politics. London: Oxford University Press.
Levinson, Stephen C. 1983. Pragmatics. Cambridge University Press.
Louw, William Ernest. 1993. Irony in the Text or Insincerity in the Writer? — The Diagnostic Potential of Semantic Prosodies. In Mona Baker, Gill Francis and Elena Tognini-Bonelli (eds.), Text and Technology. Amsterdam: John Benjamins Publishing Company. 30–50.
Machin, David, and van Leeuwen, Theo. 2007. Global Media Discourse: a critical introduction. London: Routledge.
Machin, David & Mayr Andrea. 2012. How to do critical discourse analysis. A multimodal introduction. London: Sage.
Mautner, Gerlinde. 2008. Analyzing Newspapers, Magazines and Other Print Media. In Ruth Wodak and Michał Krzyżanowski (eds.), Qualitative Discourse Analysis in the Social Sciences, 30–53. London: Macmillan Education.
Mautner, Gerlinde. 2009. Corpora and critical discourse analysis. In Paul Baker, Contemporary Approaches to Corpus Linguistics, 32–46. London: Continuum.
McEnery, Tony & Wilson Andrew. 1996. Corpus linguistics. Edinburgh: Edinburgh University Press.
McEnery, Tony, Xiao, Richard and Tono Yukio. 2006. Corpus-based Language
Studies. An Advanced Resource Book. London, New York: Routledge.
McEnery, Tony and Gavin, Brookes. 2024. Corpus linguistics and the social sciences. Corpus Linguistics and Linguistic Theory. Open access
https://www.degruyter.com/document/doi/10.1515/cllt-2024-0036/html
Mueller, John E. Presidential popularity from Truman to Johnson. The American Political Science Review, 1(64). 18-34.
Newman, David. 2002. The geopolitics of peacemaking in Israel-Palestine. Political Geography, 1(21). 629-646.
O' Halloran, Kieran. 2003. Critical Discourse Analysis and Cognitive Linguistics. Linguistics and Education, 15(1). 413–423.

Partington, Alan. 2004. Corpora and discourse, a most congruous beast. In Alan Partington, John Morley and Louann Haarman, Corpora and Discourse, 9-18. Bern: Peter Lang.
Petty, Richard E. and Cacioppo, John T. 1986. The elaboration likelihood model of persuasion. Advances in experimental social psychology, 1(19).
Popper, Karl Raimund. 2002 [1934]. The Logic of Scientific Discovery. London: Routledge.
Reisigl, Martin. 2008. Analyzing political rhetoric. In Ruth Wodak and Michal Krzyzanowski (eds.). Qualitative discourse analysis in the Social Sciences. London: Palgrave Macmillan.
Scott, Mike. 1997. PC Analysis of Key Words — And Key Key Words. System, 25(2). 233–45.
Simmel, Georg. 1908. Soziologie. Verlag von Duncker & Humblot: Leipzig.
Simpson, Paul, and Mayr Andrea. 2009. Language and Power. London: Rutledge.
Staub, Ervin. 1989. The roots of evil. The origins of genocide and other group violence. Cambridge: Cambridge University Press.
Stubbs, Michael. 1997. Whorf's children: critical comments on critical discourse analysis (CDA). In Ann Ryan & Alison Wray (eds.), Evolving models of language. Clevedon: Multilingual Matters. 55
Stubbs, Michael. 2001. Texts, Corpora, and Problems of Interpretation: A Response to Widdowson. Applied Linguistics, 22(2). 149–72.
Suurmond, Jeannine. 2005. Discourse analysis and conflict studies. Working paper 35, Netherlands Institute of International Relations Clingendael.
Teo, Peter. 2000. Racism in the News: A Critical Discourse Analysis of News Reporting in Two Australian Newspapers. Discourse & Society, 11(1). 7-49.
Thomas, William Isaac and Dorothy Swaine, Thomas. 1928. The child in America. Behavior problems and programs. New York: Alfred A. Knopf.
Titscher, Stefan, Michael Meyer, Ruth Wodak & Eva Vetter. 2000. Methods of text and discourse analysis. London: Sage.
Toolan, Michael. 1997. What is critical discourse analysis and why are people saying such terrible things about it?. Language and Literature, 6(2). 83–103.
van Dijk, A. Teun. 1988a. News as discourse. Hillsdale, NJ: Lawrence Erlbaum.
van Dijk, A. Teun. 1988b. News Analysis: Case Studies of International and National News in the Press. New York: Routledge.
van Dijk, A. Teun. 1990. News As Discourse. New York: Routledge.
van Dijk, A. Teun. 1992. Discourse and the denial of racism. Discourse & Society, 3(1). 87–118.
van Dijk, A. Teun. 1993a. Principles of critical discourse analysis. Discourse and Society, 4(2). 249–283.
van Dijk, A. Teun. 1993b. Discourse and cognition in society. In David Crowley and David Mitchell (eds.), Communication theory today. 107–126. Oxford:Pergamon Press.
van Dijk, A. Teun. 2006. Discourse and Manipulation. Discourse & Society, 17(3). 359–383.
van Dijk, A. Teun. 2015 [2001]. Critical discourse analysis. In Deborah Tannen, Deborah Schiffrin & Heidi E. Hamilton (eds.), The handbook of discourse analysis. 2nd edn, 466–485. Chichester: Blackwell.
van Leeuwen, Theo. 1995. Representing social action. Discourse Society, 6(1). 81–106.
van Leeuwen, Theo. 1996. The representation of social actors. In Carmen Rosa Caldas-Coulthard & Malcolm Coulthard (eds.), Texts and Practices. Readings in Critical Discourse Analysis, 32–70. London: Routledge.
van Leeuwen, Theo, and Ruth Wodak. 1999. Legitimizing Immigration Control: A Discourse-Historical Analysis. Discourse Studies, 1(1). 83–118.
van Leeuwen, Theo. 2005. Introducing Social Semiotics. Hove: Psychology Press.
van Leeuwen, Theo. 2008. Discourse and Practice: New Tools for Critical Discourse Analysis. Oxford University Press.
Wenden, Anita L. 2005. The politics of representation; a critical discourse analysis of an Al-Jazeera special report, International Journal of Peace Studies, 10(2). 89-112.
Widdowson, G. Henry. 1995. Discourse analysis: a critical view. Language and Literature, 4(3). 175–182.
Widdowson, G. Henry. 1998. The theory and practice of critical discourse analysis. Applied Linguistics, 19(1). 136–145.
Widdowson, G. Henry. 2000. On the Limitations of Linguistics Applied. Applied Linguistics 21(1). 3–25.

Wodak, Ruth & Michael Meyer (eds.). 2009. Methods of critical discourse analysis. London: Sage.
Wodak, Ruth. 2001a. The discourse-historical approach. In Ruth Wodak & Michael Meyer (eds.), Methods of critical discourse analysis, 63–94. London: Sage.
Wodak, Ruth. 2001b. What CDA is about—a summary of its history, important concepts and its developments. In Ruth Wodak & Michael Meyer (eds.), Methods of critical discourse analysis, 1–13. London: Sage.
Wodak, Ruth, Rudolf de Cillia, Martin Reisgl and Karin Liebhart. 2009. The discursive construction of national identity. Edinburgh: Edinburgh University Press.
Zoztman, Karin and John P. O'Regan. 2016. Critical discourse analysis and identity. In Sian Preece (Ed.). The Routledge handbook of language and identity. London:Routledge.

Specifics of language use in a polarized discourse on sex and gender: A distributional approach

Tim Feldmüller[1], **Andressa Costa**[2], **Marc Kupietz**[1]

[1]IDS Mannheim; [2]Karlsruhe Institut of Technology (KIT)

In July 2022, a controversy arose when biology doctoral student Marie-Luise Vollbrecht planned to deliver a lecture on sex, gender, and the (alleged) biological argument for only two gender categories at Humboldt University, Berlin. Prior to the planned event, Vollbrecht co-authored an article in the newspaper Die Welt that triggered debate by accusing public broadcasters in Germany of "indoctrinating" children on gender issues. The university eventually canceled her lecture, sparking widespread discussion on so-called cancel culture.

Drawing on this example, our BMBF-funded project KoKoKom investigates whether—and how—participants in polarized discourses like the one on the (non-)binarity of gender still establish a common ground. To this end, we have compiled a corpus (992,414 tokens) of social media posts and comments (YouTube, Reddit, Instagram, X, Facebook), transcripts of relevant YouTube videos and television programs, and texts from the German reference corpus DeReKo (Kupietz et al. 2018) across the period 07/2022–10/2022 containing the word Vollbrecht.

One methodological focus involves contrastive analysis of word embedding models, comparing a model trained on our project corpus with one trained on DeReKo —the largest German reference corpus (see Fankhauser & Kupietz 2017; 2019 for details). This comparison reveals, for example, that the nearest neighbors of the word *Biologin* [biologist] have shifted substantially: instead of other subdisciplinary titles like *Ökologin* [ecologist], they now refer to Vollbrecht's qualification as *Wissenschaftlerin* [scientist], reflecting a rhetorical strategy in our corpus, where *Biologin* is used to underline Vollbrecht's authority. In our poster presentation, we will introduce the KoKoKom project, describe the corpus and methodological background, and put our initial analyses and findings up for discussion.

[1] The article is available under https://www.welt.de/debatte/kommentare/plus239113451/Oeffentlich-rechtlicher-Rundfunk-Wie-ARD-und-ZDF-unsere-Kinder-indoktrinieren.html.

[2] Über Geschlecht und Gender streiten. Konflikt und Konsens als Herausforderung der Wissenschaftskommunikation [Arguing about sex and gender. Conflict and consensus as a challenge for science communication] (KoKoKom, https://kokokom.de/).

List of references

Fankhauser, Peter & Marc Kupietz. 2017. Visualizing Language Change in a Corpus of Contemporary German. In Corpus Linguistics International Conference 2017. Birmingham.

Fankhauser, Peter & Marc Kupietz. 2019. Analyzing domain specific word embeddings for a large corpus of contemporary German. In International Corpus Linguistics Conference, Cardiff, Wales, UK, July 22-26, 2019. Cardiff. https://doi.org/10.14618/IDS-PUB-9117.

Kupietz, Marc, Harald Lüngen, Paweł Kamocki & Andreas Witt. 2018. The German Reference Corpus DeReKo: New Developments – New Opportunities. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, et al. (eds.), Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018), 7-12 May 2018, Miyazaki, Japan, 8. Paris: European language resources association (ELRA).

Can ChatGPT o1 learn to analyze rhetorical moves and steps and match phrase-frames in marine science RA conclusions?

Xinyang Feng, Yingying Liu

College of Foreign Languages, Ocean University of China

Rhetorical move-step analysis has long been extensively used to examine the structural organization of academic writing. Compared with the extensive focus on introductions and abstracts, conclusion sections—where researchers synthesize findings and propose future directions—have received relatively little scholarly attention. Although recent applications of large language models (LLMs) have shown promise in automating move-step annotations, many attempts have remained at a relatively coarse-grained level. In this article, we have investigated whether ChatGPT can accurately annotate rhetorical moves and steps in the conclusion sections of marine science research articles and effectively match phrase-frames (P-frames) to these annotations. Our dataset consisted of the conclusion sections of 200 research articles in the field of marine science. As a preliminary step, we conducted a pilot study using a subset of 65 articles, which were manually annotated for move-steps based on the framework proposed by Maswana et al. (2015). We employed a few-shot learning approach, incrementally adding examples from zero-shot to five-shot settings, to examine how structured prompts could refine the model's output. The results showed substantial improvement in performance. The five-shot prompt setting was more effective in refining the model's output. Moreover, our findings indicate that ChatGPT demonstrates a promising ability to match P-frames to rhetorical move-step annotations. Our results highlight the potential of LLMs for more fine-grained rhetorical analysis and offer useful implications for researchers and educators seeking to leverage automated tools to investigate complex rhetorical structures in academic writing.

List of references

Anthony, L., & Lashkia, G. V. (2003). Mover: A machine learning tool to assist in the reading and writing of technical papers. IEEE Transactions on Professional Communication, 46(3), 185–193. https://doi.org/10.1109/tpc.2003.816789

Biber, D., Connor, U., & Upton, T. A. (2007). Discourse on the move: Using corpus analysis to describe discourse structure. John Benjamins Publishing.

Casal, J. E., & Kessler, M. (2023). Rhetorical move-step analysis. In M. Kessler & C. Polio (Eds.), Conducting genre-based research in applied linguistics (pp. 82–104). Routledge. https://doi.org/10.4324/9781003300847-7

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd-workers for text annotation tasks. Proceedings of the National Academy of Sciences, 120.

Halliday, M. A. K. (1978). Language as social semiotics. Edward Arnold.

Hyon, S. (1996). Genres in three traditions: Implications for second language teaching. TESOL Quarterly, 30(4), 693–722. https://www.jstor.org/stable/3587930

Hyland, K. (2007). Genre pedagogy: Language, literacy and L2 writing instruction. Journal of Second Language Writing, 16(3), 148–164. https://doi.org/10.1016/j.jslw.2007.07.005

Knight, S., Xie, B., Shibani, A., & Buckingham Shum, S. (2020). Are you being rhetorical? A description of rhetorical move annotation tools and open corpus of sample machine-annotated rhetorical moves. Journal of Learning Analytics, 7(3), 138–154.

Kuzman, T., Mozetič, I., & Ljubešić, N. (2023). Automatic genre identification for robust enrichment of massive text collections: Investigation of classification methods in the era of large language models. Machine Learning and Knowledge Extraction, 5(3), 1149–1175. https://doi.org/10.3390/make5030059

Lu, X., Yoon, J., & Kisselev, O. (2021). Matching phrase-frames to rhetorical moves in social science research article introductions. English for Specific Purposes, 61, 63–83.

Maswana, S., Kanamaru, T., & Tajino, A. (2015). Move analysis of research articles across five engineering fields: What they share and what they do not. Ampersand, 2, 1–11.

OpenAI. (2024). Introducing OpenAI o1 (December 5 version). https://openai.com/o1/

Paltridge, B. (1994). Genre analysis and identification of textual boundaries. Applied Linguistics, 15(3), 288–299. https://doi.org/10.1093/applin/15.3.288

Pendar, N., & Cotos, E. (2008). Automatic identification of discourse moves in scientific article introductions. In J. Tetreault, J. Burstein, & R. De Felice (Eds.), Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications (pp. 62–70). Columbus, OH: Association for Computational Linguistics. https://aclanthology.org/W08-0908

Ruiying, Y., & Allison, D. (2003). Research articles in applied linguistics: Moving from results to conclusions. English for Specific Purposes, 22(4), 365–385.
Swales, J. M. (1990). Genre analysis: English in academic and research settings. Cambridge University Press.
Swales, J. M. (2004). Research genres: Explorations and applications. Cambridge University Press.
Yu, D., Bondi, M., & Hyland, K. (2024). Can GPT-4 learn to analyse moves in research article abstracts? Applied Linguistics. https://doi.org/10.1093/applin/amae071

…And the farmer's wife provided delicious cakes: Media representations of women working in contemporary UK agriculture: A corpus linguistics approach and image analysis

Sioux Fisher

The University of Nottingham

The UK mass media plays a central and powerful role in the construction of knowledge (Carvalho, Burgess 2005; van Dijk 2014) contributing to perpetuating the gendering of occupations (Santoniccolo et al., 2023). STEM industries are crucial to driving and sustaining the UK's economic growth but persistent male dominance inhibits female participation, representing a loss of human capital for the UK economy (Hu et al. 2022).

This poster explores UK media representations of women working in contemporary UK agriculture, an industry that is attractive to women but in which women make up a small percentage of full-time employees and managers.

The Nexis archive provided UK online news articles containing references to women and men involved in UK agriculture from five national publications for an overall picture and four regional newspapers allowing regional comparison. Corpora totalling ~2,300,000 words was created. Sketch Engine is being used to examine patterns of representation around gendered pronouns and naming terms. Agency and power is being explored using a critical discourse analysis approach (Wodak and Meyer, 2016).

Initial regional media collocation analysis of terms denoting prominent women in agriculture indicates that their roles and actions have been trivialised. Additionally, generally the agricultural industry is frequently dehumanized through deletion and objectification. Future multimodal analysis will draw on images selected from salient articles during the ongoing CL analysis.

The project develops earlier research of farm wives in rural studies (Walter and Wilson 1996; Morris and Evans 2001) which identified dominant themes of women's subordination and lack of decision-making power.

The findings will be used to shape focus groups to explore whether women engaged in agricultural roles recognise their own experiences in news representations.

List of references

Carvalho, A., Burgess, J., 2005. Cultural Circuits of Climate Change in U.K. Broadsheet Newspapers, 1985–2003. Risk Analysis, 25(6), pp.1457–1469. 10.1111/j.1539-6924.2005.00692.x.

Dijk, T.A. van, 2014. Discourse and knowledge: a sociocognitive approach. New York: Cambridge University Press.

Hu, Y. et al., 2022. Gendered STEM Workforce in the United Kingdom: The Role of Gender Bias in Job Advertising [online]. UK-Canada Governments. Available at: https://committees.parliament.uk/writtenevidence/43175/pdf/ [Accessed 16 January 2025].

Santoniccolo, F., Trombetta, T., Paradiso, M. N., & Rollè, L. (2023). Gender and Media Representations: A Review of the Literature on Gender Stereotypes, Objectification and Sexualization. International Journal of Environmental Research and Public Health, 20(10), 5770. https://doi.org/10.3390/ijerph20105770

UK Parliament discourse on immigration (2007-2024) in a time of populist and post truth politics

Kenneth Fordyce, Lucia Michielin, Jessica Witte

University of Edinburgh

The past two decades have seen the increased displacement of people following sociopolitical events such as the 2008 global financial crisis, regime changes in the Middle East and North Africa[1], war in Europe, and the Coronavirus pandemic. In parallel with increased levels of migration, there has been an increase in political rhetoric and disinformation around immigration, and a rise of populist right-wing political parties portraying immigrants and asylum seekers as a threat[2,3].

This project, which investigates changes in political discourse during this turbulent period, is guided by the following research questions:

- In what ways did UK parliamentary immigration discourse change between 2007 and 2024?

- What links (if any) exist between changes in immigration discourse and specific sociopolitical events (2007-2024)?

- What patterns in immigration discourse can found in relation to individual politicians, political parties and specific governments?

A corpus of all parliamentary debates in the UK House of Commons (2007-2024) focusing on immigration or asylum was collected from the Hansard online database using data extraction techniques alongside the Hansard database's Search API. This data was imported into *OpenRefine*[4] alongside key metadata: debate date; debate title; name of politician; political party of speaker; government role (if applicable).

The dataset comprises over 13000 contributions from 374 separate debates. This poster will present the following steps in the process of data analysis:

- The refinement of the corpus data using *OpenRefine.*

- Development of an *Immigration Discourse Lexis* based on word and n-gram lists generated using *Antconc*[5].

- Statistical analysis on the relationship between the use of the *Immigration Discourse Lexis* and: (1) sociopolitical events; (2) individual language use; (3) party affiliation; (4) government affiliation (if applicable).

It will also present on the value of establishing an interdisciplinary research team involving expertise in corpus linguistics and language data extraction techniques.

List of references

(1) https://fpif.org/europes_dilemma_immigration_and_the_arab_spring/
(2) Campani, G., Fabelo Concepción, S., Rodriguez Soler, A., & Sánchez Savín, C. (2022). The rise of Donald Trump right-wing populism in the United States: Middle American radicalism and anti-immigration discourse. Societies, 12(6), 154.
(3) Boeynaems, A., Burgers, C., A. Konijn, E., & J. Steen, G. (2023). Attractive or repellent? How right-wing populist voters respond to figuratively framed anti-immigration rhetoric. Communications, 48(4), 502-522.
(4) https://openrefine.org/
(5) https://www.laurenceanthony.net/software/antconc/

Changes in linguistic markers of online self-disclosure in patients with depression

Ge Gao

University of Edinburgh, Qingdao University

For people with depression, self-disclosure of their diagnosis is a pivotal step that may alleviate feelings of isolation and encourage social support. The advent of anonymous communities on social media offers platforms for those afflicted with depression to disclose their identities and seek support with less fear of stigmatization. Despite its significance, the linguistic and psychological impacts of such disclosures require further exploration. This study examines how language use changes before and after public disclosure of a clinical depression diagnosis on online platforms.

Using data from Reddit depression-related communities, a depressive corpus was constructed, encompassing both explicit and implicit self-disclosures. It included 9,682 posts published two years before and after self-disclosure from 97 users. Linguistic Inquiry and Word Count (LIWC) was employed to measure psycholinguistic variations in emotional, cognitive, health-related, and first-person-pronoun usage. Bayesian models were applied to provide statistical explanations for differences between pre- and post-disclosure. Moreover, trend plots were utilized to exhibit temporal dynamics of attributes mentioned above.

Preliminary findings revealed a significant increase in health and mentality related terms and negative emotion words following self-disclosure. First-person singular pronouns decreased, while first-person plural pronouns increased post-disclosure, indicating reduced self-focus and heightened social engagement. By observing changing temporal trends, the findings suggested that public disclosure in anonymous settings might temporarily heighten attention to health and emotional states but also promoted social connection and psychological relief over time.

This study highlights the linguistic and psychological consequences of self-disclosure among individuals with depression. By analyzing language use before and after self-disclosure, the findings not only contribute to understanding psychological processes of individuals with depression, but also provide more linguistic evidence for depression research. Furthermore, this study also demonstrates the value of Bayesian modeling in capturing temporal linguistic changes.

Constructing the TIIB-Corpus (Corpus of Terminological Innovations in International Relations)

Mo Han, Jörn Stegmeier, Marcus Müller, Megan Fellows, Jens Steffek

Technical University of Darmstadt

Written and spoken discourses of International Relations (IR) are shaped by abstract terminology, with many emerging in academic debates or being ascribed a meaning that differs from their everyday use. It is against this background that the TIIB-Corpus is constructed for the project “Terminological Innovations in International Relations” to investigate the diffusion paths and semantic changes of key IR terms.

The corpus consists of German texts from 1976 to 2000 in three fields related to IR: 1) academic discourse – articles from selected journals, 2) political consulting by think tanks and governmental departments, and 3) political practice – plenary minutes and documents from the German Bundestag, and speeches of ministers. The collected texts are first semi-automatically cleaned and sorted using Python and Abbyy FineReader, so that the text body, footnotes and bibliography can be separated. LLMs are applied through API to correct OCR mistakes. The texts are then tokenised, lemmatised and PoS-tagged using Stanza (Peng et al. 2020), then subjected to dependency parsing using SpaCy (Honnibal & Montani 2017) and NER using flair (Akbik et al. 2018). The problems encountered in the pre-processing and annotation phases and the solutions will also be presented. The data are stored and can be browsed in CQPWeb (Hardie 2012).

The subcorpus of academic discourse has now been completed and contains 2,294 articles with a total of 13,236,245 tokens. A pilot study on "Regime" was conducted to learn 1) the efficiency of sense disambiguation of polysemic term, 2) its dependency syntactic patterns, 3) the distribution of meanings over the time span under investigation, and 4) to establish the standard workflow for studying terms.

The corpus will be made available to the academic public. Integrating three central domains of IR-related language use, it allows corpus-based research in political science and linguistic analysis of specialised languages.

List of references

Akbik, A. et al. (2018): Contextual String Embeddings for Sequence Labeling, 27th International Conference on Computational Linguistics. P. 1638-1649.

Eckart de Castilho, R., et al. (2018): INCEpTION. Corpus-based Data Science from Scratch. In: Digital Infrastructures for Research (DI4R), 9-11 October 2018, Lisbon, Portugal. Online-Ressource: http://tubiblio.ulb.tu-darmstadt.de/106982/.

Hardie, A. (2012): CQPweb. Combining Power, Flexibility and Usability in a Corpus Analysis Tool. In: International Journal of Corpus Linguistics, 17(3). P. 380-409.

Honnibal, M. & Montani, I. (2017): spaCy 2. Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Peng Q., et al. (2020): Stanza. A Python Natural Language Processing Toolkit for Many Human Languages. In: Association for Computational Linguistics (ACL) System Demonstrations. P. 101-108.

What's new in CQPweb: 2025 edition

Andrew Hardie

Lancaster University

CQPweb is an open source, browser-based concordancer, the graphical-user-interface component of the Corpus Workbench suite (https://cwb.sourceforge.io), widely used both as a standalone, locally installable program and on open servers, of which dozens now exist worldwide hosted by academic institutions and individuals with a range of research foci in terms of the type of corpus/computational linguistics facilitated and the languages, registers, and design patterns of the corpora thus made available.

2022/23 saw the addition to CQPweb of its all-time most requested feature, the ability for users to upload their own corpora, an affordance implemented using the extensibility afforded by CQPweb's system of plugins. In this contribution, I first present a post mortem on this advancement, and the limitations in the system that were revealed as a result of running it for over a year on a live server with many thousands of users (namely cqpweb.lancs.ac.uk).

I then outline the new developments put in place for version 3.4 of CQPweb which amend aspects of the system to address the issues identified. A topic of focus is the role of job scheduling in a multi-user public system where the "jobs" to be scheduled range from the interactive (e.g. compilation of frequency or collocation tables) to the long-running but high-priority (e.g. administrator operations when installing new corpora) to the low-priority (e.g. installation of user corpus data), and the intersecting needs for robustness, verboseness of error reporting, and recovery from data errors. The improvements to the user corpus upload system that the new job management system allows are illustrated – particularly in terms of the preservation of error logs for user corpora which failed to install, either due to a process crashing, or to unmanageable data formatting in the input.

Finally, other developments in CQPweb version 3.4 are summarily enumerated.

Expressions of agreement and disagreement in spoken language assessment: Evidence from the Trinity Lancaster Corpus

Ruoshi He

Lancaster University

Agreement and disagreement are essential speech acts in daily communication (Angouri & Locher, 2012). Speakers are required to master both linguistic and social knowledge to agree/disagree properly concerning interlocutors and linguistic settings, which makes it a difficult pragmatic task for L2 speakers (Gablasova & Brezina, 2018). However, despite the importance of this topic, research in this area remains limited (Bardovi-Harlig & Salsbury, 2004; Kreutel, 2007).

This study, therefore, investigated how L2 English learners at different proficiency levels agreed and disagreed in interactive spoken tasks while also exploring the influence of speaker characteristics (age, gender, and cultural background) on their linguistic choices. The study drew on data from the Trinity Lancaster Corpus, which includes 4.2 million words of transcribed L2 (candidates) and L1 (examiners) interaction in a semi-formal speaking exam. Forty-five texts (46,000 tokens) from the corpus, representing Chinese and Spanish L2 speakers, were manually coded for occurrences of agreement/disagreement. Next, automatic and semi-automatic methods were used to code the target utterances according to their linguistic renditions and pragmatic features (i.e., the strengthening and mitigating strategies used alongside agreements and disagreements). A coding system was then developed to classify different types of agreement/disagreement and examine their relationship with social variables.

Results show that higher proficiency is related to more frequent and stronger agreements and disagreements. The age and gender of the L2 speakers also played a role: female L2 speakers disagreed more frequently and adopted more diverse strategies than male speakers; older L2 speakers generally demonstrated increased values across both language functions. As for the cultural background, Spanish candidates agreed more and disagreed less than Chinese candidates, who expressed stronger disagreements with more varied strategies. These findings provide valuable insights into the development of L2 pragmatic competence across proficiency levels and also suggest practical implications for language teaching and assessment.

List of references

Angouri, J., & Locher, M. A. (2012). Theorising disagreement. Journal of pragmatics, 44(12), 1549-1553.

Bardovi-Harlig, K., & Salsbury, T. (2004). The organization of turns in the disagreements of L2 learners: A longitudinal perspective. Studying speaking to inform second language learning, 199-227.

Gablasova, D., & Brezina, V. (2018). Disagreement in L2 spoken English: From learner corpus research to corpus-based teaching materials. Learner corpus research: New perspectives and applications, 69-89.

Kreutel, K. (2007). " I'm Not Agree with You." ESL Learners' Expressions of Disagreement. tesl-ej, 11(3), n3.

Analyzing register variation in web texts through automatic segmentation

Erik Ilmari Henriksson, Antti Olavi Kanner, Veronika Maria Laippala

University of Turku

Text-linguistic analysis of registers – text varieties with shared situational characteristics and functionally related linguistic features – has greatly advanced our understanding of language variation in different contexts (Biber 1988, Biber & Conrad 2009, Biber & Egbert 2023). In the domain of online discourse, recent advances in NLP techniques such as Transformer models (Vaswani et al. 2017; Devlin et al. 2019) have enabled automatic classification of web texts into registers across various languages with near-human performance (Henriksson et al. 2024a). Despite these advances, web registers remain relatively fuzzy categories with substantial internal variation (Biber & Egbert 2020, Henriksson et al. 2024b), partly because the fundamental unit of observation – the document – often contains multiple registers.

This study introduces a novel method for analyzing register variation in web texts through classification-based register segmentation. While traditional text-linguistic register analysis treats documents as single units, we present a recursive binary segmentation approach that automatically identifies register shifts within documents without labeled segment data, using a ModernBERT classifier (Warner et al. 2024) fine-tuned on full web documents. We evaluate this method on English texts from the CORE corpus (Laippala et al. 2022) with eight main register classes. Our algorithm recursively partitions documents into register segments based on sentence boundaries, evaluating potential split points by comparing the register predictions of candidate segments and selecting those with maximally distinct predictions.

Our experimental results reveal that register segmentation produces text units with more consistent linguistic characteristics, showing lower within-register variance of linguistic features. Segments also cluster more distinctly in embedding space, with improved silhouette scores across most registers. The segment-based model outperforms the document-based model on classification tasks. Additionally, we examine register distributions within documents, revealing patterns of register shifts in online discourse. Our approach offers new insights into document-internal register variation missed by document-level analysis.

List of references

Biber, D. 1988. Variation across Speech and Writing. Cambridge University Press.
Biber, D., Conrad, S. 2019. Register, genre, and style. Cambridge University Press.
Biber, D., Egbert, J., Keller, D. 2020. 'Reconceptualizing register in a continuous situational space.' CLLT, 16(3):581-616.
Biber, D., Egbert, J. 2023. 'What is a register? Accounting for linguistic and situational variation within – and outside of – textual varieties.' Register studies 5(1):1–22.
Devlin, J., Chang, M., Lee, K., Toutanova, K. 2019. 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.' NAACL-HLT, 4171-4186.
Henriksson, E., Myntti, A., Hellström, S., Erten-Johansson, S., Eskelinen, A., Repo, L., Laippala, V. 2024a. From discrete to continuous classes: A situational analysis of multilingual web registers with LLM annotations. In NLP4DH, 308-318.
Henriksson, E., Myntti, A., Eskelinen, A., Erten-Johansson, S., Hellström, S. and Laippala, V. 2024b. 'Automatic register identification for the open web using multilingual deep learning'. arXiv preprint arXiv:2406.19892.
Laippala, V., Rönnqvist, S., Oinonen, M. et al. 2023. 'Register identification from the unrestricted open Web using the Corpus of Online Registers of English.' Lang Resources & Evaluation 57, 1045–1079.
Vaswani, A., Shazeer, N., Parmar, N. et al. 2017. 'Attention is all you need.' NIPS 30.
Warner, B, Chaffin, A., Clavié, B. et al. 2025. 'Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference.' https://arxiv.org/abs/2412.13663.

Mapping the linguistic behaviour of Korean Pop (K-Pop) fans on Twitter using corpus methods

Caitlin Hogan

Queen Mary University of London

This poster covers the provisional findings from my thesis research on the linguistic behaviour of Korean pop (K-Pop) fans on Twitter (now known as X). The field of fan studies has investigated what it means to be a fan through a variety of methods (e.g., Bennett, 2009; Jenkins, 2013; Lewis, 2003), including interviews, ethnographic methods and participant-observation, but little is known about the linguistic behaviour of fans. In other words, we know a great deal about norms of behaviour and expectations amongst different fan communities, including on Twitter, but less is known about how these norms are communicated, policed or negotiated. Similarly, little is known about whether linguistic behaviour might be a part of fan conduct. The corpus-assisted discourse analysis aims to explore what it means to be a "good" K-Pop fan, and how fans negotiate this in a fandom whose corporate components encourage competition amongst fans through their parasocial relationships (Hoffner & Bond, 2022) with K-Pop performers. The study involves a corpus of 2.4 million words obtained from 522 Twitter accounts following one of four popular K-Pop groups. This ongoing project investigates trends in function word use amongst this group and explores whether dedicated fans (those who only follow one of the four groups) differ in their language use from multi-fans (those who follow multiple groups).

Provisionally, this project finds that use of modals and pronouns differs significantly between general Twitter, multi-fans and dedicated fans. In particular, first-person singular pronouns, second-person singular pronouns and impersonal pronouns are all used significantly more by multi-fans than dedicated fans as a percentage of overall word count. This study hopes to contribute a general mapping of linguistic behaviour amongst fandom on social media, as well as how online conflict and bullying might be prevented amongst these communities.

List of references

Bennett. (2009). The Thinking Fan's Rock Band: R .E.M. fandom and negotiations of normativity in Murmurs.com.

Hoffner, C. A., & Bond, B. J. (2022). Parasocial relationships, social media, & well-being. Current Opinion in Psychology, 45, 101306. https://doi.org/10.1016/j.copsyc.2022.101306

Jenkins, H. (2013). Textual poachers: Television fans and participatory culture (Updated 20th anniversary ed). Routledge.

Lewis, L. A. (Ed.). (2003). The adoring audience: Fan culture and popular media (Transferred to digital printing). Routledge.

Estonian National Corpora 2013–2025: Methodological foundations and conctucticographic applications

Jelena Kallas, Kristina Koppel

Institute of the Estonian Langauge

This poster presents the current state of the project conducted at the Institute of the Estonian Language (IEL), focused on compiling the series of Estonian National Corpora (Estonian NC): Estonian NC 2013, 2017, 2019, 2021, 2023, and ongoing work on Estonian NC 2025. Developed by IEL in collaboration with Lexical Computing Ltd., all corpora are accessible via the Sketch Engine interface (Kilgarriff et al., 2004, 2014). The most recent Estonian NC 2023 is available, with restricted access, as a PostgreSQL database, with workflows detailed in the estnltk repository. The Estonian NC 2023, which is dependency-annotated, consists of 3.8 billion words and comprises thirteen sub-corpora. These include a biennially crawled web corpus, Wikipedia corpus, Feeds (2014–2023), and collections of old and modern literature. The corpus is semi-automatically categorized into seven genres (e.g., periodicals, forums) and 21 topics (e.g., history, food). In this poster, we will discuss the new workflow for adding sub-corpora (e.g., books from specific domains, journals published before 2016) to Estonian NC 2025 in collaboration with the National Library of Estonia's digital archive DIGAR.

Using the diverse functionalities of Sketch Engine, the Estonian NC series has been utilized in various lexicographic projects and has served as a primary resource for detecting collocations (Kallas et al., 2015) for the Estonian Combined Dictionary, the largest dictionary for modern Estonian (Koppel et al., 2019). Using Estonian NC 2023 as an example, we will showcase the use of dependency relations for the automatic detection of collocations (previously, we employed a Part-of-Speech pattern approach; cf. Uhrig, Proisl, 2012) and phrasal constructions. For this purpose, we utilize the Word Sketch functionality of Sketch Engine. The Estonian Sketch Grammar has been specifically adapted to enable corpus analysis based on morpho-syntactic annotation. Additionally, pre-defined queries are executed directly from the relational database to capture specific morpho-syntactic co-occurrences.

List of references

DIGAR. The National Library of Estonia's digital archive. https://www.digar.ee/arhiiv/en

estnltk. *estnltk-workflows: enc_workflows*. GitHub, 2025, https://github.com/estnltk/estnltk-workflows/tree/master/enc_workflows.

Kilgarriff, Adam; Rychlý, Pavel; Smrz, Pavel; Tugwell, David 2004. The Sketch Engine. – Proceedings of the 11th EURALEX International Congress, 105–115.

Kilgarriff, Adam; Baisa, Vít; Bušta, Jan; Jakubíček, Miloš; Kovář, Vojtěch; Michelfeit, Jan; Rychlý, Pavel; Suchomel, Vít 2014. The Sketch Engine: Ten years on. – Lexicography, 1, 7–36. https://doi.org/10.1007/s40607-014-0009-9

Koppel, Kristina; Tavast, Arvi; Langemets, Margit; Kallas, Jelena 2019. Aggregating dictionaries into the language portal Sõnaveeb: issues with and without a solution. In: Kosem, I., Zingano Kuhn, T., Correia, M., Ferreria, J. P., Jansen, M., Pereira, I., Kallas, J., Jakubíček, M., Krek, S. & Tiberius, C. (Ed.). Proceedings of the eLex 2019 conference. 1-3 October 2019, Sintra, Portugal.. (434−452). Brno: Lexical Computing CZ, s.r.o.

Kallas, Jelena; Kilgarriff, Adam; Koppel, Kristina; Kudritski, Elgar; Langemets, Margit; Michelfeit, Jan; Tuulik, Maria; Viks, Ülle 2015. Automatic generation of the Estonian Collocations Dictionary database. Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference, 11-13 August 2015, Herstmonceux Castle, United Kingdom. Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd, 1−20.

Uhrig, Peter; Proisl, Thomas 2012. Less hay, more needles – using dependency-annotated corpora to provide lexicographers with more accurate lists of collocation candidates. Lexicographica, vol. 28, no. 2012, pp. 141-180. https://doi.org/10.1515/lexi.2012-0009

‘The takeaways doctors want to cancel’, ‘miracle foods’, and ‘culinary exotica’ for the ‘fearless eater’ - representations of Chinese foodways in the UK press ‘Post-Covid’

Ursula Kania, Lorna Ryan

University of Liverpool

In May 2023, the World Health Organisation (WHO) declared that Covid-19 was no longer considered a ‘public health emergency of international concern’. However, the repercussions of this global crisis can still be felt on many levels. One socio-cultural legacy is the continued stigmatisation of (predominantly East & Southeast) Asians, triggered by the outbreak's initial association with a 'wet market' in China, and exacerbated by media coverage drawing on longstanding stereotypes of (alleged) Chinese foodways (e.g., consuming bats; Kania, 2023). As the 2022-23 UK Home Office report shows, the 'new normal' includes approx. 33,600 hate crimes targeting Asians per year.

Given the media's role in shaping public perceptions, this study – situated within corpus-assisted critical discourse analysis (e.g., Baker & McEnery, 2015) - aims to analyse UK press coverage of Chinese food culture 'post-Covid', with a particular focus on evaluating if/to which degree negative stereotypes of (alleged) Chinese foodways still persist, and if/how representations may have been affected by Covid-19.

The corpus, collected using LexisNexis, consists of 1,483 relevant articles from 12 UK broadsheets, tabloids, and regional newspapers, published in the 12-month-period following the WHO announcement (i.e., May 2023 - April 2024), totalling 1,524,719 words. Qualitative coding is currently underway in NVivo, and quantitative analyses (keywords&collocates) will be conducted shortly, using AntConc (Anthony, 2024).

Initial qualitative results suggest that while coverage of (British-)Chinese food culture is very multi-faceted - ranging from the popularity (and ‘unhealthiness’) of takeaways to alleged ‘miracle food[s]’ - (‘traditional’/’authentic’) Chinese food is still often construed as the exotic (and thus potentially inherently dangerous) other.

Overall, the results of the study will provide insights into the representation of Chinese foodways in the British press ‘post-Covid’, including the continuing use and hence normalisation of problematic stereotypes which may be (re)mobilised in harmful ways (not only) in times of crisis.

List of references

Anthony, L. (2024). AntConc (Version 4.2.4) [Computer software]. Retrieved from https://www.laurenceanthony.net/software

Baker, P. and McEnery, T. (eds.) (2015). Corpora and Discourse: Integrating Discourse and Corpora. London: Palgrave.

Kania, U. (2023). ‘They really eat anything don't they?’: Pronoun use in COVID-19-related Anti-Asian racism. In L. Paterson (ed.). The Routledge Handbook of Pronouns. London: Routledge, pp. pp. 333-349.

Using ChatGPT to clean large corpora: The case of the UK 'knife crime' British newspaper corpus

Aleksei Konshin

University of Glasgow

This paper explores the potential of generative AI, presenting a Python-based custom script designed by ChatGPT to refine a 66.9-million-word newspaper corpus on the topic of 'knife crime'. Although existing studies predominantly assess capabilities for text analysis in corpus-based research (Curry et al., 2023; Zappavigna, 2023), the application of generative AI in corpus cleaning remains underexplored. To bridge this gap, a Python script was developed to tackle a pervasive issue common to corpora derived from LexisNexis data: sampling errors caused by the use of semantically ambiguous search terms when retrieving news articles (Kantner et al., 2011).

Based on a 122-word query specifying Python as the programming language and Tkinter for the user interface, ChatGPT developed the 'Cleaner Corpus Caretaker' tool. The query outlined the script's core function: extracting filenames from column A of an .xlsx document and removing corresponding .txt files from corpus folders. Subsequent queries expanded the code from 77 to 201 lines, introducing a log file for tracking removed documents and an 'undo action' feature.

To address sampling errors, concordance lines and filenames, copied from the AntConc (Anthony, 2024) results window, were inserted and examined in an Excel file. These lines were generated by words and phrases found in publications deemed irrelevant. For instance, the term *'axed'* appeared in news about mass layoffs, captured due to the search term *'axe'* employed in LexisNexis. After retaining concordance lines corresponding to irrelevant files, the 'Cleaner Corpus Caretaker' excluded documents from the corpus folder using the remaining filenames in the Excel file. By applying this procedure across 11 categories of false positives (e.g., cooking recipes, advertisements), the 'knife crime' corpus was reduced from 66.9 to 29.4 million words.

Overall, this case demonstrates how generative AI can enable non-specialists to develop custom data cleaning solutions for corpora, without requiring prior coding knowledge.

List of references

1. Anthony, L. (2024). AntConc (4.3.1.) [Software]. Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software/AntConc
2. Curry, N., Baker, P., & Brookes, G. (2023). Generative AI for corpus approaches to discourse studies: A critical evaluation of ChatGPT. Applied Corpus Linguistics, 4(1), 100082. https://doi.org/10.1016/j.acorp.2023.100082
3. Kantner, C., Kutter, A., Hildebrandt, A., & Püttcher, M. (2011). How to get rid of the noise in the corpus : cleaning large samples of digital newspaper texts. International Relations Online Working Paper Series, 2011/2. https://www.uni-stuttgart.de/soz/ib/forschung/IRWorkingPapers/IROWP_Series_2011_2_Kantner_Kutter_Analysis_Newspaper_Texts.pdf
4. Zappavigna, M. (2023). Hack your corpus analysis: How AI can assist corpus linguists deal with messy social media data. Applied Corpus Linguistics, 3(3), 100067. https://doi.org/10.1016/j.acorp.2023.100067

Linguistic portrayals of England's national football teams: A corpus-based analysis of gendered discourses in UK newspapers

Yipei Kou[1,2]

[1]Cardiff University; [2]Hebei Sport University

Since the Lionesses' Euro 2022 win and reaching the 2023 World Cup final, the UK media's changing approach to women's football, with increased coverage and less sexualising language, indicates a major shift in sports portrayal. However, achieving equal visibility to men's football remains challenging.

This study investigates how UK newspapers linguistically represent the England men's and women's national football teams. A diachronic specialised corpus was constructed from news articles covering six FIFA World Cups (2014, 2018, 2022 for men and 2015, 2019, 2023 for women). Articles were sourced from Nexis Advance, including 18 influential British national newspapers covering each tournament period to capture peak media attention. The search terms were carefully developed to maximise relevance and minimise noise, incorporating key terms such as "England," "Three Lions," and "Lionesses," alongside context-specific phrases like "FIFA" and "World Cup." These terms were refined through iterative testing to ensure precision and to exclude unrelated mentions of other sports or events.

The study employs an integrated methodology that combines corpus linguistics (CL) and critical discourse analysis (CDA). The study first examines the men's and women's sub-corpora separately through word frequency and collocation analyses to identify the semantic domains and discourse prosodies present in each corpus. Building on these insights, a gender comparison is conducted using keyness analysis to highlight distinctive patterns, alongside diachronic comparisons to track changes over time. The analysis also includes Social Actor Analysis to examine role allocation, focusing on how the media represents the England teams as active agents or passive recipients. These methods allow for a detailed investigation of both linguistic patterns and the broader ideology they reveal.

This study contributes to bridging the gap between linguistics, gender studies, and sports media research. It sheds light on societal norms, power structures, and cultural values surrounding gender in sports media.

Register variation in Latin

Hanna-Mari Kupari

University of Turku

My presentation explores register variation in Latin across millennia in regards to grammatical features. Registers, as defined by Biber and Conrad (2009), are text varieties produced in a specific communicative situations. They can be studied by their linguistic features, such as part-of-speech (POS) tags and other structural characteristics, that are functionally tied to a communicative situation.

Previous research on register variation in Latin is limited. Chaudhuri et al. (2019) investigated linguistic variation between two genres—poetry and prose. Their methodology focused on features using string searches, such as the frequency of cum as a subordinating conjunction and superlative forms containing *-issim-*.

Building on parsed text corpora with machine-readable grammatical syntactic annotations, i.e. treebanks, Hudspeth et al. (2024), classified texts into eleven categories (e.g., epic, treatise) and incorporated this as metadata to a 887 K token corpus. I extend this corpus manually marking register types (Biber & Conrad, 2005) as metadata (e.g. personal letter, religious) to the CIRCSE treebank and the Classical LASLA[1] (20 K and 1.8 M tokens) and Medieval texts of the Corpus Corporum[2].

As a method, I use key feature analysis KFA (Egbert & Biber, 2023), a quantitative approach, to investigate if in a data based exploratory approach using all Universal Dependiencies guideline annotated POS and morphological features predictions (Kupari et al., 2024) can distinguish register varieties. In KFA texts in two registers with respect to a set of functionally motivated grammatical features. Frequency counts from Hudspeth et al. (2024) dataset motivated by Chaudhuri et al., (2019). demonstrate relative clauses using *cum* more frequent in satire than epic and superlative forms in letters than legal texts (three and five times respectively).

Examining registers using the KFA method in UD datasets enables a more nuanced analysis than word-based searches and deepens our understanding of Latin usage in diverse contextual settings.

List of references

1 https://github.com/CIRCSE/LASLA/tree/main

2 https://mlat.uzh.ch/browser?path=/

Biber, Douglas, and Susan Conrad. "Register Variation: A Corpus Approach." In The Handbook of Discourse Analysis, edited by Deborah Schiffrin, Deborah Tannen, and Heidi E. Hamilton, 2005.

Biber, Douglas, and Susan Conrad. "Registers, Genres, and Styles: Fundamental Varieties of Language." In Register, Genre, and Style, Cambridge: Cambridge University Press, 2009. 1–28. Print. Cambridge Textbooks in Linguistics.

Chaudhuri, Pramit, Tathagata Dasgupta, Joseph P. Dexter, and Krithika Iyer. "A Small Set of Stylometric Features Differentiates Latin Prose and Verse." Digital Scholarship in the Humanities 34, no. 4 (December 2019): 716–729.

Egbert, Jesse, and Douglas Biber. "Key Feature Analysis: A Simple, Yet Powerful Method for Comparing Text Varieties." In Corpora 18, no. 1 (2023): 121–133.

Hudspeth, Marisa, Brendan O'Connor, and Laure Thompson. "Latin Treebanks in Review: An Evaluation of Morphological Tagging Across Time." In Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024), 203–218, Hybrid in Bangkok, Thailand, and online. Association for Computational Linguistics, 2024.

Kupari, Hanna-Mari Kristiina, Erik Henriksson, Veronika Laippala, and Jenna Kanerva. "Improving Latin Dependency Parsing by Combining Treebanks and Predictions." In Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities, 216–228, Miami, USA. Association for Computational Linguistics, 2024.

Representing homelessness: A corpus-assisted discourse analysis of BBC News UK

Sayuri Kusama, Gavin Brookes

Lancaster University

Homelessness presents a major challenge in the UK, where more than 390,000 people experiencing it. However, research suggests that public awareness surrounding homelessness is often limited and fails to comprehend its scale and complexities. Mass media is one of the primary forces shaping public perceptions of social issues yet, importantly, their representations are always partial, resulting from journalistic decisions around what to foreground and background. Such decisions, and the kinds of 'framings' that result, are profoundly shaped by the news actors' commercial and ideological aims (Fairclough, 1995; Fowler, 1991; van Dijk, 1988).

Despite the widespread nature of this problem, few studies have adopted systematic linguistic approaches to explore how homelessness is represented by the media in the UK context (Gómez-Jiménez & Bartley, 2023; Parnell, 2023, 2024). The current research aims to add insight to this small body of work by examining, for the first time, how homelessness is framed by the popular news website, *BBC News UK*. The data analysed consists of 358 news articles published by *BBC News UK Online* in 2022-2023 (157,854 words). Using *LancsBox X*, a keyword-driven framing analysis is carried out, based on close-reading of 100 concordances for the top 52 keywords (Baker, 2023; Entman, 1993; Goffman, 1986). The keywords were obtained by comparing the specialised corpus against BE21 with the significance (log-likelihood) test.

The research demonstrates that the BBC predominantly foregrounds responses to and experiences of homelessness rather than its causes, mirroring a long-standing trend amongst commercial and partisan news media. Our analysis identified housing and polycrisis discourses that could help interpret the issue from structural perspectives, as well as a 'nuisance' discourse that may stigmatise people experiencing homelessness. Notably, political actors in the central and devolved governments are far less salient than local councils and charity organisations in the corpus.

List of references

Baker, P. (2023). Using Corpora in Discourse Analysis. Bloomsbury Publishing Plc.

Baker, P., Brookes, G., Atanasova, D., & Flint, S. W. (2020). Changing frames of obesity in the UK press 2008–2017. Social Science & Medicine, 264, 113403. https://doi.org/10.1016/j.socscimed.2020.113403

Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. Journal of Communication, 43(4), 51–58. https://doi.org/10.1111/j.1460-2466.1993.tb01304.x

Fairclough, N. (1995). Media discourse. E. Arnold.

Fowler, R. (1991). Language in the News: Discourse and Ideology in the Press. Taylor & Francis Group.

Goffman, E. (1986). Frame analysis: An essay on the organization of experience (Northeastern University Press ed). Northeastern University Press.

Gómez-Jiménez, E. M., & Bartley, L. V. (2023). 'Rising Number of Homeless is the Legacy of Tory Failure': Discoursal Changes and Transitivity Patterns in the Representation of Homelessness in The Guardian and Daily Mail from 2000 to 2018. Applied Linguistics, 44(4), 771–790. https://doi.org/10.1093/applin/amac079

Parnell, T. (2023). 'A tide of homeless, drug-addicted and mentally ill people': Representing homeless people in MailOnline content. Journal of Corpora and Discourse Studies, 6, Article 1. https://doi.org/10.18573/jcads.97

Parnell, T. (2024). 'Homeless' and the cost-of-living crisis in The Guardian and MailOnline: A corpus-assisted analysis. In T. Parnell, T. Van Hout, & D. Del Fante (Eds.), Critical Approaches to Polycrisis: Discourses of Conflict, Migration, Risk, and Climate, pp. 125–146. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-76966-5_6

van Dijk, T. A. (1988). News As Discourse. Routledge.

Corpus-based linguistic analysis of Muslim hate speech in South Korea: Focusing on the Mosque Construction Incident in Daegu

Jun Lee[1], Minkyu Sung[1], Jinsan An[2], Kilim Nam[1]

[1]Yonsei University; [2]Kyungpook National University

This study aims to analyze hate speech targeting Muslims, specifically focusing on the "Mosque Construction Incident" in Daegu, South Korea, in 2020, and to construct a semantically annotated 'Muslim Hate Speech Corpus.' Additionally, this research seeks to establish a semantic annotation framework for domain-specific texts and develop training data for ethical AI.

According to Park (2023), although South Korea is increasingly becoming a multicultural society with 2.2 million immigrants, hatred and discrimination against them remain significant social challenges (Mutuma, 2015; Park, 2023). However, societal awareness of racism remains limited, and research addressing this issue has been inadequate. In contrast, international research has made substantial progress in analyzing semantic domains for hate speech and building datasets to study immigrants hate speech(Baker et al., 2013; Yue et al., 2024; Al Mandhari et al., 2024).

This research comprises three parts. First, hate speech triggered by the "Mosque Construction Incident" will be collected from social media to build a Muslim Hate Speech Corpus (one million words), examining the characteristics of hate expressions in user-generated texts. Second, by comparing and analyzing hate speech annotation systems, such as 'HD (Assaults on Human Dignity)' and 'CV (Calls for Violence)' proposed by Kennedy et al. (2018), the study will develop a hate speech annotation framework tailored to Korean contexts. This will include annotation schemes for Korean-specific MWEs and unregistered words in Korean hate speech. Third, the research will construct a semantically annotated hate speech corpus to perform corpus linguistic analyses, such as keyword extraction, key semantic category analysis, and collocation analysis, alongside sociolinguistic analyses.

This study is significant for its investigation into the shared characteristics and distinctive features of Korean hate speech. The findings are expected to contribute to policymaking aimed at fostering immigrant integration into society and the development of unbiased generative AI.

List of references

Al Mandhari, S., El-Haj, M., & Rayson, P. (2024). Is it Offensive or Abusive? An Empirical Study of Hateful Language Detection of Arabic Social Media Texts.

Baker, P. (2013). Discourse Analysis and Media Attitudes: The Representation of Islam in the British Press. Cambridge University Press.

Kennedy, B., Atari, M., Davani, A. M., Yeh, L., Omrani, A., Kim, Y., ... & Dehghani, M. (2018). The gab hate corpus: A collection of 27k posts annotated for hate speech. PsyArXiv. July,18.

Park, S. (2023). The future of hate and discrimination: Policy and legislative alternatives. National Assembly Futures Institute.

Ruteere, M. (2013). Report of the Special Rapporteur on contemporary forms of racism, racial discrimination, xenophobia and related intolerance. Addendum: visit to Spain.

Sanguinetti, M., Poletto, F., Bosco, C., Patti, V., & Stranisci, M. (2018, May). An italian twitter corpus of hate speech against immigrants. In Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018).

Yue, T., Shi, X., Mao, R., Hu, Z., & Cambria, E. (2024, May). SarcNet: A multilingual multimodal sarcasm detection dataset. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) (pp. 14325-14335).

The lexical bundling of EFL teachers' questions in higher education classrooms in the Indonesian context

Agustina Lestary

University of Limerick

Many studies have suggested that teacher's questions hold significant role in classroom interaction as questions can help teacher measure students' background knowledge and evaluate students' understanding. Questions are particularly important in language classroom because it can promote students' engagement and help students to produce the target language. Despite the growing number of research on teachers' questions in different settings, there are still limited studies applying Corpus Linguistics in analysing teachers' questions in the classroom, particularly in Indonesian settings. The focus of this research is identifying the most frequent lexical bundling in teachers' questions in different classroom modes and their pedagogical purpose. Classroom modes are the terms introduced by Walsh (2006) who claimed that single context aiming for one pedagogical objective only does not exist in any classroom, including language classroom. Thus, Walsh proposed four contexts (or modes) that could portray the fluidity of the contexts in classroom. Around 38,000 words are collected from 7 classrooms of English Language Education Departments in different universities in Indonesia. NVivo is used to classify the classroom modes and Sketch Engine is used to identify the most frequent lexical bundling. Scaffolding is the most frequent type of questions identified in classroom context and material modes. Despite having the same function, teachers used different lexical bundling for scaffolding in different modes. The teacher mostly used 'what else' in classroom context mode as a cue for the students to supply more responses. On the other hand, the teacher tried to measure students understanding on the materials by asking 'any other opinion' or 'other opinion. The use of these lexical bundling as questions had allowed more students to engage in the discussion.

Insights from the Languaging Crises project: Corpus-assisted approach to public crisis discourse during COVID-19

Hanna Limatius, **Jenni Räikkönen**

University of Helsinki

The Languaging Crises (LanCris) project focuses on the linguistic and discursive features of government crisis communication in the United Kingdom and the United States during three historically impactful pandemics: the late 1800's tuberculosis pandemic, the 1918 influenza pandemic ("Spanish flu"), and the COVID-19 pandemic. We study authorities' public crisis communication (via e.g., news, parliamentary debates, social media), as well as citizens' responses to this communication. The focus of the project is on texts published in English.

The poster introduces two work-in-progress studies conducted in the LanCris project that utilize methods of corpus-assisted discourse analysis (Partington, Duguid & Taylor 2013). The first study explores in-group and out-group discourse (Wodak 2008) in citizens' responses to authorities' social media communication during COVID-19. The data for this study come from the Instagram account of the Centers for Disease Control and Prevention (CDC) and include comments from 216 posts published in 2020 and 2021 (~430,000 tokens). The study illustrates how citizens not only label authorities as an out-group, but also construct group boundaries among themselves based on different views of the crisis.

The second study focuses on expressions of epistemicity by the U.S. President Donald Trump and British Prime Minister Boris Johnson in their speeches related to the COVID-19 pandemic in 2020. Honesty and openness are necessary to build trust during a crisis (Seeger 2006), but political leaders may be tempted to hide any uncertainty. The corpus used in the study is comprised of 78 speeches overall (~134,000 tokens) collected from the White House archives and the British government's website.

The two studies illustrate different perspectives to the crisis: those of citizens and leaders. In our conclusion, we argue that both perspectives are necessary to evaluate and develop current crisis communication practices. Inclusive and clear crisis communication is key in managing on-going and future crises.

List of references

Partington, A., A. Duguid, and C. Taylor. 2013. Patterns and Meanings in Discourse: Theory and Practice in Corpus-Assisted Discourse Studies (CADS) [Studies in Corpus Linguistics 55]. John Benjamins.

Seeger, M. W. 2006. Best Practices in Crisis Communication: An Expert Panel Process. Journal of Applied Communication Research 34(3), 232–244. https://doi.org/10.1080/00909880600769944.

Wodak, R. 2008. 'Us' and 'Them': Inclusion and Exclusion – Discrimination via Discourse. In Identity, Belonging and Migration, eds. G. Delanty, R. Wodak, and P. Jones, 54–77. Liverpool University Press.

Investigating connectivity change in the constructional network

Hewei Liu, Chadi Ben Youssef, Martin Hilpert

University of Neuchâtel

This poster presents work in progress that uses corpora to investigate change in constructional networks. Diachronic Construction Grammar (Noël 2007; Traugott & Trousdale 2013; Hilpert 2021) offers a theoretical framework for understanding language change through processes like constructionalisation and constructional shifts. These processes can be conceptualised as 'network changes', involving nodes and their connectivity within the constructional network (Sommerer & Smirnova 2020). Connectivity refers to associative linkages between linguistic units, including sequential, lexical, paradigmatic, taxonomic, and lexico-grammatical relationships (Diessel 2019). Despite foundational research, many open questions remain regarding the dynamics of these various types of connectivity. This study presents a new corpus-based and data-driven approach to analysing diachronic changes in connectivity within a large constructional network. The goal is to explore how different types of connectivity develop over time. Our approach is inspired by distributional semantics (Turney & Pantel 2010) and behavioural profile analysis (Gries & Divjak 2009). We use the Corpus of Historical American English (COHA; Davies 2010). A vector space for 2,000 lexical items — including nouns, verbs, adjectives, and adverbs — is analysed across decades. For each item, concordance lines are extracted to examine co-occurrence frequencies across three levels: lexical collocations, partially specified syntactic patterns, and schematic constructions. This process generates a co-occurrence frequency matrix, producing unique constructional profiles for each lexical item over time. The poster will focus on the methodological steps of the approach, its implementation, and expected analytical results. By leveraging corpus data for the purpose of network analysis, we aim to strengthen the empirical foundations of Diachronic Construction Grammar. Our long-term goal is to establish principles that characterise connectivity changes, contributing to the refinement of Diachronic Construction Grammar as a general theory of language change and offering new insights into historical linguistics through network dynamics.

List of references

Davies, Mark. 2010. The Corpus of Historical American English (COHA): 400+ Million Words, 1810–2009, available online at http://corpus.byu.edu/coha.

Diessel, Holger. 2019. The Grammar Network: How Linguistic Structure Is Shaped by Language Use. Cambridge: Cambridge University Press.

Gries, Stefan Th. and Dagmar S. Divjak. 2009. Behavioral profiles: a corpus-based approach towards cognitive semantic analysis. In Vyvyan Evans & Stephanie S. Pourcel (eds.), New directions in cognitive linguistics, 57-75. Amsterdam & Philadelphia: John Benjamins.

Hilpert, Martin. 2021. Ten Lectures on Diachronic Construction Grammar. Leiden: Brill.

Noël, Dirk. 2007. Diachronic construction grammar and grammaticalization theory. Functions of Language 14: 177–202.

Sommerer, Lotte and Elena Smirnova (eds.) 2020. Nodes and Networks in Diachronic Construction Grammar. Amsterdam: John Benjamins.

Traugott, Elizabeth C. and Graeme Trousdale. 2013. Constructionalization and Constructional Changes. Oxford: Oxford University Press.

Turney, Peter and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. Journal of Artificial Intelligence Research 37, 141-188.

Register variation in written Chinese legal texts: A novel multi-dimensional analysis

Yang Liu

School of Liberal Arts, Anhui University

Previous quantitative studies on the Chinese legal register have generally examined only a few varieties at a time, focusing on a limited set of linguistic characteristics. Thus, this study explores register variation in written Chinese legal texts based on self-built large-scale Chinese Corpus of Legal Texts and novel Multi-Dimensional Analysis framework. It aims to address the following three questions: (1) In a corpus consisting of the Sub-corpus of Chinese Legislative Texts and the Sub-corpus of Chinese Judicial Texts, what dimensions of variation are observable? (2) Based on these dimensions, how distinct are Chinese legislative texts from Chinese judicial texts? (3) What distinctive linguistic features of the Chinese legal register can be observed and what underlying reasons contribute to its formation? Distinctive characteristics of the Chinese legal register will be considered when selecting linguistic features, while machine learning algorithms, such as the Random Forest Model and K-Means clustering, will be employed to verify the reasonableness of the selected linguistic features and count these features. Quantitative multi-dimensional data derived from factor analysis will be used to describe and compare the linguistic features of two types of Chinese legal texts (legislative texts and judicial texts). Qualitative investigation will focus on functional interpretation based on contrastive data of co-occurring linguistic features, and the reasons behind the formation of distinctive features of the Chinese legal register. The contributions of this study are twofold. On the one hand, it can enrich the field of register studies to a certain extent, especially from the perspective of Chinese legal register. On the other hand, it can help to validate, improve, and expand the initial Multi-Dimensional Analysis framework proposed by Biber (1988) from the perspective of Chinese legal register and contribute somewhat to the future integration of Multi-Dimensional Analysis research with machine learning and natural language processing techniques.

List of references

Biber, D. (1984). A model of textual relations within the written and spoken modes. University of Southern California Doctoral Dissertation.

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

Biber, D., & Egbert, J. (2018). Register variation online. Cambridge: Cambridge University Press.

Ervin-Tripp, S. (1972). On sociolinguistic rules: Alternation and co-occurrence. In Gumperz, J., & Hymes, D. (Eds.), Directions in sociolinguistics (pp. 213–250). Cambridge: Basil Blackwell.

Goźdź-Roszkowski, S. (2011). Patterns of linguistic variation in American legal English: A corpus-based study. Frankfurt am Main: Peter Lang.

Huang, Y., & Ren, W. (2020). A novel multidimensional analysis of writing styles of editorials from China Daily and The New York Times. Lingua, 235, 102781.

Huang, Y., & Sang, Z. (2024). Linguistic variation in supreme court oral arguments by legal professionals: A novel multi-dimensional analysis. Discourse Studies, 14614456231221075.

Jang, S. C. (1998). Dimensions of spoken and written Taiwanese: A corpus-based register study. University of Hawai'i at Manoa Doctoral Dissertation.

Pan, Q. Y. (1991). Investigations of Chinese legal register. Kunming: Yunnan People's Publishing House.

Pang, S. Z., & Wang, K. F. (2023). Investigating the diachronic change of literary register features based on Chinese translated and original corpora. Journal of Foreign Languages, 46(6), 77–78.

Ren, C., & Lu, X. (2021). A multi-dimensional analysis of the management's discussion and analysis narratives in Chinese and American corporate annual reports. English for Specific Purposes, 62, 84–99.

Tai, Q., & Rao G. (2021). A study on the measurement and classification of Chinese stylistic features. In Proceedings of the 20th Chinese National Conference on Computational Linguistics (pp. 398–412). Huhhot, China.

Thompson, P., Hunston, S., Murakami, A., & Vajn, D. (2017). Multi-dimensional analysis, text constellations, and interdisciplinary discourse. International Journal of Corpus Linguistics, 22(2), 153–186.

Tiersma, P. M. (1999). Legal language. Chicago: University of Chicago Press.

Xu, J. J., & Li, J. L. (2022). A study of register change in written Chinese over the last century. Foreign Languages and Their Teaching, (4), 76–86+147–148.

Zhang, Z. S. (2017). Dimensions of variation in written Chinese. New York: Routledge.

Can a quantitative linguistic model, combined with machine learning, classify and analyze the English variants? A case study of English variants in Olympic Journalism

Felix Mao

Rye Country Day School

The majority of existing research on global English variants has heavily relied on qualitative approaches. In recent years, the emergence of advanced language models and machine learning techniques have created new opportunities for stylometric text style analysis. However, the empirical viability of these approaches in quantitative stylometric research for English variants has not yet been fully tested. This study illuminates the capabilities and limitations of corpus-driven quantitative linguistic models in their ability to classify and analyze the stylometric differences of English variants in Olympic Journalism. A corpus was constructed by sourcing 1810 Olympic news documents in American English, Chinese English, and Spanish English from 2020 to 2023. A set of models was developed by incorporating 164 linguistic features with the high-level representations of text mining methods and advanced language models (TF-IDF, BERT embeddings, and GPT embeddings), and three machine learning algorithms were applied (Naïve Bayes, XGBoost, and Support Vector Machines (SVM)). There are two notable findings from this study. First, the quantitative linguistic models performed well in classifying the English variants. In particular, the GPT embedding incorporated with the advanced linguistic features and the SVM machine learning algorithm was the most effective model, even outperforming the current state-of-the-art GPT embedding (F1-score 97.2 vs. 96.9). Second, as part of the model assessment, a list of statistically significant linguistic features were identified, and a follow-up analysis reveals how these features from linguistic perspective contributed to stylometric difference. The research results also revealed the most influential linguistic features that differentiate American, Chinese, and Spanish English, both quantitatively and qualitatively. By integrating the corpus-driven quantitative models and qualitative linguistic feature analysis, this study demonstrated the enhanced capability to classify and analyze English variants and evidenced potential opportunities to expand such an integrated approach in stylometric research and broader linguistic research.

Mapping ideological positions through social media profiles: A case study of X users (WITHDRAWN)

Luciana Nogueirol Lobo Marcondes[1], Maria Claudia Nunes Delfino[2]

[1]Faculdade de Tecnologia de São Paulo; [2]PUC-SP - Pontifícia Universidade Católica de São Paulo

User profile texts on social media can reveal the political orientation of the users, yet these texts have received little attention in research. This study addresses this gap by analyzing the discourses used in user profiles to align themselves with specific political stances. A corpus of profiles from users on the social media platform X was collected, focusing on individuals who expressed their views on the Supreme Federal Court (STF) of Brazil during the COVID-19 pandemic. In many countries, Supreme Courts played a significant role in pandemic management, addressing constitutional challenges and balancing public health measures with individual freedoms. In Brazil, the STF faced accusations of constitutional overreach, favoritism toward foreign and corporate interests, and curtailment of personal liberties. This study collected and analyzed profile texts in Portuguese from users who collectively posted a total of 27358 texts referring to the Supreme Course. The data were analyzed using Multi-Dimensional Lexical Analysis (Berber Sardinha, 2019, 2020; Berber Sardinha & Fitzsimmons-Doolan, 2024; Fitzsimmons-Doolan, 2019, 2023), which is geared to the identification of discourse constructs in corpora using a Multi-Dimensional framework (cf. Biber, 1988). Six dimensions of ideological positioning were identified, each with positive and negative poles. For example, Dimension 1 captures a contrast between conservative nationalism, emphasizing national sovereignty and traditional values, and ideologies advocating individual freedom and criticizing judicial interventionism. Overall, the analysis reveals how users' self-presentation in their profiles encodes broader discursive disputes regarding social order, the relationship between individuals and the State, and the tension between tradition and change. These findings highlight the potential of user profile texts as a lens for understanding the ideological alignment of social media users and their engagement with political institutions.

List of references

The authors acknowledge the financial support of the following organizations: Programa CAPES – Programa Estratégico Emergencial de Prevenção e Combate a Surtos, Endemias, Epidemias e Pandemias

Número do Processo: 88887.703941/2022-00; National Council for Scientific and Technological Development (CNPq), Grants # 310140/2021-8, 403130/2024-7, 444019/2024-3

Metadiscourse features in food promotional genres: Being informed or persuaded?

Maria Melissourgou

University of the Aegean

During a product promotional campaign, marketing professionals try to highlight its competitive advantage over other similar products, but also to inform about its key features in case the product is not known in the target country (generic advertisement). Even though the most easily recognizable promotional genre is *advertising*, new and often mixed / hybrid genres, are emerging.

Raising genre awareness in such a specialised context can empower ESP (English for Specific Purposes) educators and educational material writers in guiding marketing professionals. ESP material is scarce and knowledge on genres among educators is usually implicit rather than explicit (Handford, 2010; Rea, 2010). Seen from a different angle, readers / consumers are harder to manipulate when they have a conscious understanding of the purpose of a text, especially when this purpose is difficult to recognise, as in the case of mixed genres.

The *Agri-Food corpus,* a specialised genre-based corpus (794,811 tokens), has been designed and constructed with this aim in mind (Melissourgou & Frantzi, 2021, 2025). It consists of texts which promote agri-food products in industry magazines, trade fair brochures and company websites. The method for text identification has been based on Systemic Functional Grammar theoretical principles. It takes into account the functional purpose (Martin, 1985) and the register variables, namely field, tenor and mode (Halliday, 1978). Three, out of the five genres included in this corpus, namely the *Advertisement, Web Product Presentation* and *Generic Product Infomercial* are explored in this paper. Based on the model of *Metadiscourse* (Hyland 2005), the three genres are investigated for interactional resources (hedges, boosters, attitude markers, engagement markers and self-mentions). The paper identifies clear differences between presenting information about a product and intending to persuade readers in order to consume.

List of references

Halliday, M.A.K. (1978). Language as social semiotic: The social interpretation of language and meaning. London: Edward Arnold.

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

Hyland, K. (2005). Metadiscourse: Exploring interaction in writing. London, UK: Continuum.

Martin, J.R. (1985). Factual writing: Exploring and challenging social reality. Geelong, Deakin University Press (Reprinted 1989.Oxford, Oxford University Press).

Melissourgou, M. N. & Frantzi, K. T. (2021). Generic patterns in promotional texts: the case of the Greek agri-food business website. In the 12th International Conference on Corpus Linguistics (CILC2021) abstract book (p. 114-115). Virtual, University of Murcia, Spain.

Melissourgou, M. N. & Frantzi, K. T. (2025). Promotional Genres in the Agri-Food Corpus. In Selected papers of the 56th Linguistics Colloquium. Linguistik International. Frankfurt am Main: Peter Lang.

Rea, C. (2010). Getting on with Corpus Compilation: from Theory to Practice. ESP World, 1(27), Volume 9.

Exploring the effectiveness of DDL in mandatory undergraduate English education in Japan

Aika Miura

Rikkyo University

This poster presentation explores the implementation and the effectiveness of Data-Driven Learning (DDL) in a mandatory Reading & Writing course for 23 first-year undergraduates at a Japanese private university during the 2024 academic year. Targeting students with CEFR A2-B1 proficiency, the course focused on developing academic reading and writing skills through strategies such as skimming, annotating, drafting, and revising, and vocabulary building.

Building on previous research (Miura & Satake, 2024), DDL was introduced through a combination of self-study and in-class activities, utilising various corpus tools, including SKELL and Sketch Engine. Activities encouraged students to analyse word meanings, collocations, and contextual usage of target vocabulary.

The study explores the effectiveness of DDL through a series of assignments culminating in a 200-word descriptive paragraph about students' favourite dishes. These tasks included:

(i) Planning the paragraph: Generating descriptive adjectives about the dish's ingredients, associated feelings, and cultural or personal significance

(ii) Conducting DDL activities: Using the English Web Corpus, enTenTen21, in Sketch Engine, to analyse adjectives from a textbook model paragraph (e.g., "citrusy," "fresh," "appetizing" to describe a dish called "Ceviche") and students' chosen words

(iii) Drafting and revising: Incorporating reasons for liking the dish and detailed sensory descriptions with DDL-informed adjectives chosen by the students

Despite pedagogical and time constraints such as a unified syllabus and a previously assigned textbook, which are commonly encountered in English Language Teaching (ELT) in tertiary education in Japan, the students expressed high satisfaction with DDL. However, an analysis of their completed paragraphs indicated that its impact on practical writing skills was limited, as previously observed by Miura and Satake (2024). This preliminary study underscores the potential of DDL in fostering lexical awareness while highlighting the need for further refinement to enhance its application in academic writing for lower-level English learners in Japan.

List of references

Miura, A., & Satake, Y. (2024). Implementation of Data-Driven Learning (DDL) in a course for English Reading and Writing. Journal of Multilingual Pedagogy and Practice, 4, 1–17.

The emergence of the JB-X DM-Y construction

Yuzo Morishita

Momoyama Gakuin University

Hirose (1991) implies that the construction (hereafter, JB-X DM-Y construction), as shown in (1a), is a reduction from a general complex sentence, as shown in (1b).

(1) a. Just because John is rich doesn't mean that he is happy.

b. Just because John is rich, it doesn't mean that he is happy. (Hirose 1991:25)

In addition, Bender & Kathol (2001:18) also argue that the subject is omitted in the construction. This is thought to be because, as Quirk et al. (1985:1047) noted, there is no other construction in which the subject is a *because*-clause (cf. Hilpert 2007:30).

However, the results of a diachronic survey using the Corpus of Contemporary Historical American English (COHA) contradict the assertions of these previous studies. The mono-clausal one, shown in (1a), is always freuent than the bi-clausal one, shown in (1b).

In this study, I argue that the construction illustrated in (1a) did not emerge through the reduction of (1b), but rather through the analogy of the construction illustrated in (2).

(2) a. That doesn't mean your work is amateurish. (COHA 1999)

b. Eddie is a college boy too, but that doesn't mean he knows enough to stay away from the pool hall. (COHA 2008)

In fact, I observed the similarity of frequency trends between JB-X DM-Y construction and *that doesn't mean* construction in (2) from COHA.

These findings suggest that the JB-X DM-Y construction emerged not from syntactic reduction but from analogy with constructions such as (2). This aligns with theoretical frameworks such as Hilpert's (2017) diachronic construction grammar. By integrating frequency-based evidence with theoretical insights, this study contributes to a deeper understanding of the mechanisms driving constructional change and analogical innovation.

List of references

Hirose, Y. (1991) On a certain nominal use of because-clause: Just because because-clause can substitute for that-clause does not mean that this is always possible. English Linguistics 8, 16–33.

Bender, E. & A. Kathol. (2001) Constructional effects of 'just because...doesn't mean...'. Proceedings of Annual Meeting of the Berkeley Linguistics Society 27(1), 13–25.

Quirk, R., S. Greenbaum, G. Leech, & J. Svartvik. (1985) A comprehensive grammar of the English language. Longman.

Hilpert, M. (2021) Ten lectures on diachronic construction grammar. Brill Academic Pub.

Home on the Range with genAI: A corpus-assisted analysis of frontier and foundation metaphors in the discursive conceptualization of Large Language Models

Mike Murphy

Carleton University

Large language models (LLMs) have been hailed as a breakthrough in genAI research that will have beneficial applications in many domains (Shone, 2024). However, skeptics have concerns about the LLM project in terms of, e.g., the ethics of training data harvesting, the environmental effects of these systems, and the risk that the vast amounts of money and expertise they require could lead to economic benefits for an exclusive few (Bender et al., 2021). Drawing on Critical Metaphor Analysis (Charteris-Black, 2004) and Corpus Assisted Discourse Studies (Baker, 2023), this study seeks to answer the following research questions:

- Since the 2022 appearance of ChatGPT 3.5, what sorts of metaphors have been most prominently used in public written texts from corporate AI labs to conceptualize LLMs?
- What do the particular metaphors employed by corporate AI labs suggest about which aspects of LLM technology they wish to "highlight" and "hide" (Lakoff & Johnson, 1980)?

A pilot for a larger, doctoral project, this study involved the construction of a roughly 300,000-token specialized corpus. The corpus consists of written texts from the websites of OpenAI and Anthropic, including announcements, product descriptions, and policy statements. Key LLM metaphors were identified through collocational analysis, with "LLM" and its equivalents as the node, followed by concordance analysis, and frequency counts for metaphor-related lexical items.

The dominant metaphors for conceptualizing LLMs were found to be the frontier and the foundation. These exist in tension, the first evoking the frontier myth in American history--and the violent, acquisitive, colonial ideologies it entails (Grandin, 2019); and the second suggesting that this potentially threatening new technology is in fact safe, solid, and reliable. This dyad of conceptual metaphors functions ideologically to characterize LLM technology as an infinite growth opportunity while simultaneously downplaying the risks and costs associated with LLM development.

List of references

Baker, P. (2023). Using corpora in discourse analysis (2nd ed.). Bloomsbury Academic.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623. https://doi.org/10.1145/3442188.3445922

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

Grandin, G. (2019). The end of the myth: From the frontier to the border wall in the mind of America. Metropolitan Books.

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

Shone, O. (2024, October 9). 5 key features and benefits of large language models. The Microsoft Cloud Blog. https://www.microsoft.com/en-us/microsoft-cloud/blog/2024/10/09/5-key-features-and-benefits-of-large-language-models/

Developing a corpus of Finnish literary texts for children

Tapio Nojonen, Filip Ginter, Veronika Laippala, Jenna Kanerva, Mikko-Jussi Laakso

University of Turku

This poster presents an ongoing project developing a corpus of literary texts written for children aged 5 and up in Finnish. Corpora containing such data have been created for a variety of different languages (e.g. Korochkina et al. 2024; Li et al. 2023; Schroeder et al. 2015), but currently none exist for Finnish. This greatly hampers our ability to analyze and understand the linguistic characteristics of texts targeted at children of these ages. One of the main applications our project foresees is to use the data in the development of language models to recognize, create and model educational material for different age groups.

To decide which books to include in the corpus, we looked at the e-shops of the most popular Finnish book retailers and divided books into three different age groups: 7-8, 9-12, and 13+. These age groups were determined by the guidelines for Finnish in The National Core Curriculum for Basic Education (Opetushallitus 2014). We also checked if the list was reasonable by seeing if it contained the most popular books mentioned in a national study called Lukuklaani, which surveyed the reading habits of students in Finnish basic education (Grünthal et al 2019).

To obtain the text from the books, we did the following: first scan the books as pictures, then extract text with Google Cloud Services, and finally syntax parse with the Trankit dependency syntax parser (Nguyen et al 2021). Protected by the EU Text and Data Mining exception, we have scanned 240+ books as an initial dataset. It contains over 8 million words and around 207,000 unique lemmas, of which ~0.8% appear in more than half the books and ~66% only in a single book. In the poster, we will also show results for age-specific subcorpora, such as what age word types are first encountered at.

List of references

References

Grünthal, S., Hiidenmaa, P., Routarinne, S., Tainio, & Aaltonen Lotta-Sofia (2019). ”Lukuklaani-tutkimus”. url: https://blogs.helsinki.fi/lukuklaani/

Korochkina, Maria et al. (2024). ”The Children and Young People’s Books Lexicon (CYP-LEX): A large-scale lexical database of books read by children and young people in the United Kingdom”. Quarterly Journal of Experimental Psychology. Publisher: SAGE Publications, s. 17470218241229694. issn: 1747-0218. doi: 10.1177/17470218241229694. url: https://doi.org/10.1177/17470218241229694

Li, Luan et al. (2023). ”CCLOWW: A grade-level Chinese children’s lexicon of written words”. Behavior Research Methods 55.4, s. 1874–1889. issn: 1554-3528. doi: 10.3758/s13428-022-01890-9. url: https://doi.org/10.3758/s13428-022-01890-9

Nguyen, Minh Van et al. (2021). ”Trankit: A Light-Weight Transformer-based Toolkit for Multilingual Natural Language Processing”. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations.

Schroeder, Sascha et al. (2015). ”childLex: a lexical database of German read by children”. Behavior Research Methods 47.4, s. 1085–1094. issn: 1554-3528. doi: 10.3758/s13428-014-0528-1. url: https://doi.org/10.3758/s13428-014-0528-1

Opetushallitus (2014). ”Perusopetuksen opetussuunnitelman perusteet 2014”. url: https://www.oph.fi/fi/koulutus-ja-tutkinnot/perusopetuksen-opetussuunnitelman-perusteet

Using language models for linguistic analysis: The discourse function of initial vs. final dependent structures

Naoki Otani[1], Ryo Nagata[2], Hiroya Takamura[3], Yoshifumi Kawasaki[4]

[1]Tokyo University of Foreign Studies / Lancaster University; [2]Konan University; [3]The National Institute of Advanced Industrial Science and Technology; [4]The University of Tokyo

This presentation explores the potential of word and clause vectors to complement and supplement traditional, time-consuming multifactorial corpus analyses. Since vectors are generated based on contextual information without predefined grammatical rules (Vaswani et al. 2017; Devlin et al. 2019), they are particularly suited to capture characteristics that emerge from actual language use (Authors 2024a, 2024b). This presentation applies word and clause vectors to test a hypothesis on discourse coherence (Halliday and Hassan 1976).

According to Thompson (1985), sentence-initial dependent clauses, as in (1), often serve global or discourse functions, while sentence final-dependent clauses, as in (2), tend to serve local functions:

(1) To cool, place the loaf on a wire rack.

(2) Place the loaf on a wire rack to cool.

Here, (1) implies to state a 'problem' within the context of expectations raised by the preceding discourse, while, in (2), the dependent clause functions more locally, adding the purpose to the main clause. To evaluate Thompson's hypothesis, this study analyses vector similarities among 1) sentence-initial or sentence-final dependent clauses, 2) main clauses, and 3) preceding contexts (2 sentences). The methodology includes:

1) Validating vector similarity as a measure of discourse coherence by comparing examples from coherent and less coherent texts, before the main corpus experiment

2) Extracting some hundred examples of initial and final dependent structures (e.g., *to*-purpose clauses) respectively from large corpora (e.g., *Corpus of Contemporary American English*)

3) Using language models (LMs) to generate vector representations for dependent clauses, main clauses, and preceding contexts

4) Measuring vector similarities to assess the degree of coherence and evaluate the hypothesis

If sentence-initial dependent clauses exhibit higher vector similarity with preceding contexts than sentence-final clauses, Thompson's hypothesis will be supported by AI-based analysis. This study highlights the potential of vector representations as a tool for advancing corpus-based discourse analysis.

List of references

Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova (2019) "BERT: Pre-training of deep bidirectional transformers for language understanding," Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1.

Halliday, M. A. K. and Hasan, Ruqaiya (1976) Cohesion in English. London: Longman.

Thompson, Sandra A. (1985) "Grammar and written discourse: Initial vs. final purpose clauses in English," Text 5: 55-84.

Vaswani, Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin (2017) "Attention is all you need," Advances in Neural Information Processing Systems 30 (NIPS 2017).

Discourse markers in intercultural collaborative exchanges during a transnational telecollaborative project

Samed Yasin Ozturk

Mus Alparslan University

Discourse markers play a crucial role in facilitating communication by ensuring mutual understanding between participants (Aijmer, 2004). These are lexical items that are essential in creating textual coherence and expressing interpersonal and interactional feelings and views (Carter & McCarthy, 2006). Although discourse markers have been studied extensively in learner language, there has been limited research based on data from intercultural collaborative exchanges among learners from different countries and L1 backgrounds, where discourse markers can potentially play a crucial role in maintaining interactional fluency and fostering mutual understanding.

Accordingly, this corpus-based study examines the use of discourse markers in a six-week telecollaborative project that incorporated Global Citizenship Education (GCE), Sustainable Development Goals (SDGs), and Intercultural Communicative Competence (ICC) among student teachers from Turkey, Germany, and Palestine. Drawing on Byram's (2021) ICC framework, the study compiled a corpus of interactions from recordings of weekly online team meetings, collaborative tasks, and semi-structured interviews. All the interactions in the recordings were transcribed verbatim for corpus analysis. The corpus contained approximately 150,000 words, and the analysis focused on five commonly studied markers: 'so,' like,' you know,' 'I mean,' and 'well,' examining their frequency and discourse functions within this cross-cultural context. For statistical comparisons, the log-likelihood statistic and the Mann-Whitney U test were used. Initial findings reveal distinctive patterns in the use of these markers, with Turkish students demonstrating a tendency toward textual over interactional functions, pointing out potential areas for targeted pedagogical intervention. The results will be discussed with reference to the literature on discourse markers in learner and native speaker English.

List of references

Aijmer, K. (2004). Pragmatic markers in spoken interlanguage. Nordic Journal of English Studies, 3(1), 173-190.

Byram, M. (2021). Teaching and assessing intercultural communicative competence (Revised edition). Multilingual Matters.

Carter, R., & McCarthy, M. (2006). Cambridge grammar of English: A comprehensive guide to spoken and written English grammar and usage. Cambridge: Cambridge University Press.

'I falter where I firmly trod': Charting metaphors of bereavement in elegy

Eva Rothenberg

University of Birmingham

Conceptual metaphors of motion (LIFE IS A JOURNEY, DEATH IS DEPARTURE) underpin our relationship with life and death. Cultural discourse surrounding grief also suggests the underlying conceptual metaphor GRIEF IS A JOURNEY. Elegy presents a compelling avenue for exploring bereavement because it operates in the personal and public domains of grief and mourning, respectively. There have been limited investigations of conceptual metaphor in poetry. This study explores metaphors of movement and stasis in Alfred Tennyson's *In Memoriam, A. H. H.*, and how their distribution expresses the psychological contours of the speaker's grieving process.

The text was composed over 17 years, making it an interesting diachronic case study. It comprises 133 poems (19,154 tokens), which served as small-scale corpus. Examining the 100 most frequent content words in context using AntConc informed my search for conceptual metaphors. Approximately 20% were related to mortality, such as death, soul, time, and change. Many also evidenced motion, such as come, go, leave, thro' (i.e. through), out, and away. The corpus was then split in half, with 67 poems in one sub-corpus and 68 in the other. A keyword analysis ($p < 0.01$) was performed for each sub-corpus to explore how themes of grief and healing manifest diachronically in the text.

Using an adapted version of Steen et al.'s (2010) MIPVU procedure, I manually coded the poem for metaphor-relatedness pertaining to speaker movement and stasis. A rolling average of each category was calculated, and graphs produced to illustrate metaphorically-dense clusters. My analysis highlights the nonlinearity of mourning. Emotional progress is tempered by strong moments of grief that induce an almost-paralytic state. Embodied metaphors of healing are also shaped by theological conceptions of an afterlife. This study explores how corpus tools can inform coding practices for poetry, which poses unique challenges to established metaphor identification procedures.

List of references

Fainsilber, L. and Ortony, A. (1987) 'Metaphorical Uses of Language in the Expression of Emotions', Metaphor and Symbol, 2(4), pp. 239–250.
Lakoff, G. and Johnson, M. (1980) Metaphors We Live By. Chicago: University of Chicago Press.
Lakoff, G. and Turner, M. (1989) More than Cool Reason: a Field Guide to Poetic Metaphor. Chicago: University of Chicago Press.
Littlemore, J., et al. (2014) 'An Investigation into Metaphor Use at Different Levels of Second Language Writing', Applied Linguistics, 35(2), pp. 117–144.
Littlemore, J. & Turner, S. (2020) 'Metaphors in communication about pregnancy loss', Metaphor and the Social World, 10(1), pp. 45–75.
Parkes, C. M. (1971) 'Psycho-social transitions: A field for study', Social Science & Medicine, 5(2), pp. 101–115.
Steen, G. S., et al. (2010) A Method of Linguistic Metaphor Identification. Amsterdam: John Benjamins.
Tennyson, A. (1908) 'In Memoriam, A. H. H.' in E. Gray (ed.), In Memoriam (2004), Norton Critical Edition, 2nd edn. New York: Norton.

Building a longitudinal learner corpus of underrepresented EFL genres

Pamela Saavedra-Jeldres

Universidad Católica de Temuco

This poster presents an ongoing project to build the longitudinal Chilean corpus of learner English, spanning elementary to advanced proficiency levels (A2 to C1 of the CEFR). The study is conducted within a 5-year undergraduate English language teacher education (ELTE) programme that incorporates EMI and CLIL approaches. Graduates of this programme will teach EFL in secondary schools, emphasizing the importance of developing their writing skills during university studies. Despite the growing availability of learner corpora, there remains a need for datasets including underrepresented EFL genres such as emails, letters, reports, and reviews (McEnery et al., 2019).

The corpus will collect data at 12 points over two years, tracking the same 200 learners from group years 1 to 5. Each student will provide 12 texts written under exam conditions (handwritten form) to ensure AI-free writing. Texts are transcribed following guidelines by Brenchley & Durrant.

This corpus will provide valuable insights into vocabulary acquisition and instruction, as vocabulary is critical in language learning and writing performance (Laufer & Nation, 1995). Lexical features, including density, diversity, and sophistication, will be analysed using Text Inspector, TAALED (Kyle & Crossley, 2015), and TAALES (Kyle & Crossley, 2018). Descriptive and inferential statistics will identify developmental trends, addressing the following research question: *What developmental trends in lexical complexity—density, diversity, and sophistication—emerge in a longitudinal learner corpus* mediated by cohort, proficiency level, and genre*?*

Data collection for the first two semesters is complete, and transcription is underway. Findings will inform L2 writing instruction in underrepresented EFL genres, shedding light on students' vocabulary size, depth, and developmental trajectories across cohorts and proficiency levels.

List of references

Brenchely & Durrant. Growth in Grammar Project Transcription Manual - Stage One. University of Exeter. https://phildurrant.net/annotation-manual-2/

Kyle, K., & Crossley A., S. (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

Kyle, K., & Crossley, S. A. (2015). Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application. TESOL Quarterly, 49(4), 757-786. https://doi.org/10.1002/tesq.194

Laufer, B., & Nation, P. (1995). Vocabulary Size and Use: Lexical Richness in L2 Written Production. Applied Linguistics, 16(3), 307-322. https://doi.org/10.1093/applin/16.3.307

McEnery, T., Brezina, V., Gablasova, D. & Banerjee, J. (2019). Corpus linguistics, learner corpora, and SLA: Employing technology to analyse language use. Annual Review of Applied Linguistics, 39, 74-92. https://doi.org/10.1017/S0267190519000096

Politicizing public health: The discourses around public health organizations

Tatiana Schmitz de Almeida Lopes[1,2,3], Maria Claudia Nunes Delfino[1], Fernanda Peixoto coelho[1,2,3]

[1]Pontifical Catholic University of São Paulo; [2]Faculdade de Tecnologia de Praia Grande; [3]Universidade Metropolitana de Santos - UNIMES

Before the COVID-19 pandemic, public health organizations like WHO, CDC, ANVISA, ECDC, and PAHO operated largely unnoticed by the general public, focusing on health policy and disease control. However, the pandemic brought these agencies into the spotlight, especially in heavily impacted regions like the US and Brazil. Increased visibility was accompanied by politicization and criticism of their management strategies. Actions such as mask mandates, vaccination policies, and lockdowns, implemented with health authority support, became ideologically charged issues. This led to conflicting representations of these organizations, influenced by diverse political and ideological discourses, ranging from trust to denialism.

Despite this heightened visibility, there has been no comprehensive study examining how health organizations were represented on social media. Addressing this gap, the current research analyzes Brazilian tweets in Portuguese using a corpus of approximately 88,000 messages mentioning major health organizations. The corpus was created with the Python tool “Twarc,” using search terms like “pandemic,” “WHO,” “ANVISA,” and “Covid-19.” Results were saved in a structured JSON file, cleaned to remove duplicates, and prepared for analysis.

The study employs Lexical Multidimensional Analysis (LMA), an extension of Multidimensional Analysis. Unlike traditional approaches that describe register variation, LMA identifies patterns in discourses, ideologies, and themes.(Berber Sardinha, 2017, 2021, 2023; Berber Sardinha & Fitzsimmons-Doolan, in prep.; Clarke et al., 2021, 2022; Fitzsimmons-Doolan, 2014, 2019, 2023) Factors identified encapsulate how health organizations were portrayed during the pandemic. This method helps reveal public understanding, perceptions, and challenges in communicating health guidelines, aiming to foster better strategies to support public trust and defend health as a fundamental right, aligning with the Universal Declaration of Human Rights.

List of references

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). London: Bloomsbury.

Berber Sardinha, T. (2021). Discourse of academia from a multi-dimensional perspective. In E. Friginal & J. Hardy (Eds.), The Routledge Handbook of Corpus Approaches to Discourse Analysis (pp. 298-318). Abingdon: Routledge.

Berber Sardinha, T., & Fitzsimmons-Doolan, S. (2024). Lexical Multidimensional Analysis. Cambridge: Cambridge University Press.

Biber, D. (1988). Variation Across Speech and Writing. Cambridge: Cambridge University Press.

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. (2023). 21st century ideological discourses about US migrant education that transcend registers. Corpora, 18(2), 143-173.

A corpus-based approach to cultural heritage and climate change: Investigating terminological usage in policy documents

Elisa Squadrito[1], **Monica Monachini**[2]

[1]University of Macerata; [2]CNR-ILC

While climate change is extensively researched, its impacts on Cultural Heritage have only recently emerged as a distinct investigative concern (European Commission, 2022). Intergovernmental bodies, such as the European Union, have been at the forefront in supporting frontier research in Heritage and Climate sciences. Yet, despite growing policy interest, a lack of standardized terminology hinders interdisciplinary collaboration between Heritage practitioners and Climate scientists (Simpson et al., 2022).

This project, conducted in collaboration with CLARIN (De Jong et al., 2022) and E-RIHS (Cano Díaz et al., 2019) Research Infrastructures, addresses these challenges by investigating the specialized language used in policies addressing Cultural Heritage and Climate Change. It explores three central questions: (1) Which terms recur in policy documents related to Cultural Heritage and Climate Change? (2) How do these terms compare to those related to Climate Science? (3) In what ways can terminological misalignments impede effective policymaking.

To address these questions, a specialized corpus of grey literature from supranational projects that explore the intersection of Climate Change and Cultural Heritage is being compiled. Texts selected for inclusion are represented by policy documents—such as Green and White Papers, technical and assessment reports—addressing both climate and heritage issues.

Corpus-based methods (Sinclair, 1991; McEnery, 2019; McEnery & Brookes, 2024) will guide the extraction and analysis of keywords, multi-word expressions, and phraseology. A reference corpus of IPCC assessment reports (Agrawala, 1998) will provide a comparative baseline. E-RIHS experts will ultimately validate candidate terms and offer conceptual feedback, shaping the final glossary. The project's outputs will include the corpus and a computational terminological resource, both accessible through CLARIN, facilitating more consistent communication and informed decision-making among linguists, policy experts, and other stakeholders. By highlighting terminological overlaps and discrepancies, this research offers a novel contribution to bridging disciplinary divides in Cultural Heritage and Climate Change policy.

List of references

Agrawala, S. (1998). Context and early origins of the IPCC. Climatic Change, 39, 605–620.

Bowker, L., & Pearson, J. (2002). Working with specialized language: A practical guide to using corpora. Routledge.

Cano Díaz, E., et al. (2019). The European Research Infrastructure for Heritage Science (E-RIHS): An infrastructure for an interdisciplinary scientific domain. [Publication details not provided].

European Commission: Directorate-General for Education, Youth, Sport and Culture. (2022). Strengthening cultural heritage resilience for climate change – where the European Green Deal meets cultural heritage. Publications Office of the European Union. https://data.europa.eu/doi/10.2766/44688

de Jong, F., et al. (2022). Language matters: The European Research Infrastructure CLARIN, today and tomorrow. In D. Fišer & A. Witt (Eds.), CLARIN: The infrastructure for language resources (Vol. 1, pp. 31–58). De Gruyter. https://doi.org/10.1515/9783110767377-002

McEnery, T. (2019). Corpus linguistics. Edinburgh University Press.

McEnery, T., & Brookes, G. (2024). Corpus linguistics and the social sciences. Corpus Linguistics and Linguistic Theory, 20(3), 591–613. https://doi.org/10.1515/cllt-2024-0036

Simpson, N. P., & Orr, S. A. (2022). Impacts, vulnerability, and understanding risks from climate change to culture and heritage. ICOMOS & ICSM CHC.

Sinclair, J. (Ed.). (1991). Corpus, concordance, collocation. Oxford University Press.

Exploring regional variation and written representation of Chinese dialects

Qi Su

University of Birmingham

Language variation and change in Chinese languages remain critical yet underexplored areas, particularly in comparing regional variations of Mandarin, Cantonese, Shanghainese, and other dialects. This study investigates linguistic features and regional language variations across geographically distinct regions, such as Beijing, Shanghai, Guangzhou, and northeastern China. It aims to provide insights into modern Chinese language variation and change.

A central focus of this research is the comparison of linguistic features in spoken and written data, including social media platforms like Weibo. Specific research questions include whether linguistic features found in spoken data also appear in written forms and how these features differ across two different datasets. The study will further explore the linguistic pattern in Mandarin to represent dialectal expressions that lack official written forms. For example, Cantonese speakers may use characters with similar sounds in Mandarin to approximate Cantonese expressions, such as using '唔' (meaning 'not'), adapting them for phonetic and semantic purposes in informal online contexts.

Mandarin: 我 不 知 (wo bù zhī)

1SG NEG know

I don’t know

Cantonese: 我 唔 知 (ngo5 m4 zi1)

1SG NEG know

I don’t know

The research will employ a corpus-based methodology, combining qualitative and quantitative analyses to identify patterns of variation. Data collection includes naturally occurring spoken interactions and written social media content from diverse regions. The study considers sociolinguistic factors, such as gender, age, economic status and regional identity, to contextualize findings and explore how these factors influence language use.

This research hopes to fill a critical gap by moving beyond single-dialect analyses, offering a comprehensive perspective on the regional variation of Chinese languages. Future work will refine the analytical framework and expand the dataset to explore additional features and their implications for modern Chinese linguistics.

Gain and loss framing in racial hate crime communication: A corpus-based approach

Suphanut Sukkasem

University of Leicester

In the UK, the official figures of reported crimes are believed to significantly underrepresent the actual number of incidents. In combating the underreporting of racial hate crimes, law enforcement agencies use online platforms to communicate and provide information to the public. The way information is framed emphasising either benefits of performing a particular action (gain framing) or negative consequences of not performing an action (loss framing) could have different impacts on the recipients' decision-making. This concept is rooted in Prospect Theory (Kahneman and Tversky, 1984) which explains how people evaluate potential gains and losses differently depending on how choices or outcomes are presented. However, there is a lack of knowledge about the application of gain and loss framing in hate crime communication.

This study explores whether there is evidence of gain and loss framing in the information about racial hate crime communicated to the public and potential victims, as well as the extent to which framing may influence victims' decision to report hate crime. To achieve this, the study will combine corpus linguistics methods and an experiment. A specialised corpus, the Racial Hate Crime Corpus, is under construction and will consist mainly of webpages from government and law enforcement agencies in England and Wales. It would include a sub-corpus from non-governmental organisation websites for comparison. The corpus will be analysed for gain and loss framing language. Subsequently, participants from East and Southeast Asian background will be shown framed messages extracted from the corpus and asked to rate how encouraging or discouraging the message seem on the Likert scale. The study aims to provide insight into how framed messages influence decision-making in hate crime reporting and may contribute to developing guidelines for effective communication strategies that the law enforcement agencies can use to encourage crime reporting among victims and the public.

List of references

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. American psychologist, 39(4), 341.

Pezzella, F. S., Fetzer, M. D., & Keller, T. (2019). The dark figure of hate crime underreporting. American Behavioral Scientist, 0002764218823844.

Protection Approaches (2024). East and Southeast Asian Communities' Experiences of Hate Crime in the UK. Available at: https://protectionapproaches.org/esea-hate-crime-report.

Measuring the impartiality of UK parliamentary research services using corpus linguistics

Xinmei Sun

Lancaster University

The UK House of Commons Library aims to provide impartial information services for MPs and MPs' staff regardless of their political affiliation, thereby helping them "scrutinise legislation, prepare for debates, develop policies" (Commons Library, n.d.). Their briefing reports are described as the "first port of call" and an aid to decision-making by MPs (Kenny et al., 2017).

Given the crucial part that the Commons Library plays in legal and political decision-making, it is critical to assess whether they are indeed impartial as mandated. This project develops and tests a way of measuring impartiality in texts using automatic annotation and corpus tools. It then applies this method to determine the impartiality of a corpus of briefing reports on the topic of immigration published between 1999-2024.

Impartiality is defined as the presentation of information without privileging a particular party or expressing personal opinions. Drawing from cognitive linguistics, it is approached through framing and epistemic positioning.

This poster presents the first stage in this project and assesses semantic frame annotation tools trained using FrameNet (Ruppenhofer et al., 2016). I consider: 1) the reliability of the automatic annotation of semantic frames and frame elements when compared against manual annotation in a sample of sentences from the briefing reports; 2) how the tools' performance can be enhanced with limited manual annotation resources; and 3) the potential benefits of using semantic frames as the analytical unit (as opposed to words). Using frequency measure and concordancing, I consider: 4) the distribution of competing frames on a policy issue; 5) the treatment of sources with different political affiliations; 6) the use of overt evaluations. This poster will discuss the implications of the findings for impartiality and how corpus linguistics can contribute to the development of a paradigm for monitoring impartiality.

List of references

House of Commons Library. (n.d.). About us. Retrieved October 7, 2024, from https://commonslibrary.parliament.uk/about-us/

Kenny, C., Rose, D.C., Hobbs, A, Tyler, C. & Blackstock, J. (2017). The Role of Research in the UK Parliament, Volume One. https://www.parliament.uk/globalassets/documents/post/The-Role-of-Research-in-the-UK-Parliament.pdf

Ruppenhofer, J., Ellsworth, M., Petruck, M., Johnson, C.R., Baker, C.F. & Scheffczyk, J. (2016). FrameNet II: Extended Theory and Practice. https://framenet2.icsi.berkeley.edu/docs/r1.7/book.pdf

Can large language models understand sarcastic hate speech? An experimental study on detecting sarcastic hate speech targeting Muslims in Korean

Minkyu Sung[1], Jun Lee[1], Jinsan An[2], Kilim Nam[1]

[1]YONSEI University; [2]Kyungpook National University

The evaluation of Large Language Models (LLMs) in natural language generation and understanding has been extensively conducted across various domains, with some studies suggesting that LLMs exhibit language capabilities comparable to humans. However, experimental investigations reveal that LLMs still underperform in domain-specific or genre-specific identification and discourse analysis tasks (Curry et al., 2024; Uchida, 2023). Similarly, interpreting sarcastic or hateful expressions, which require subjective judgment within context, remains a challenge for LLMs (Turban and Kruschwitz, 2022). This study assesses the performance of LLMs in analyzing sarcastic hate speech, focusing on their ability to detect and interpret unethical language. By comparing the judgment results of human experts with those of generative AI, this research evaluates the performance and limitations of LLMs in analyzing sarcasm within hate speech.

The research methodology consists of the following steps: (1) using the YouTube Data API to collect 30,000 comments containing Muslim hate speech, (2) constructing a dataset of human-annotated sarcastic hate speech (approximately 1,000 comments) through usage-based analysis, and (3) conducting evaluation experiments with LLMs following prompt development. For instance, a comment like, “Building a street of pork restaurants in front of a mosque would make them happy,” qualifies as sarcastic hate speech, leveraging external knowledge about Islam's prohibition of pork to convey hate.

A pilot analysis of 200 sarcastic hate speech comments using ChatGPT’s general model revealed an error rate of 60–75% in detecting sarcastic expressions, significantly higher than for explicit hate speech. These findings highlight the limitations of LLMs in detecting context-dependent sarcastic hate speech. This study compares the performance of LLMs in detecting explicit hate speech and sarcastic hate speech, emphasizing the essential role of contextual understanding in developing ethical AI. Furthermore, it classifies subtypes of sarcastic hate speech and discusses the reasons for errors from a linguistic perspective.

List of references

Curry, N., Baker, P., & Brookes, G. (2024). Generative AI for corpus approaches to discourse studies: A critical evaluation of ChatGPT. Applied Corpus Linguistics, 4(1), 100082.

Turban, C., & Kruschwitz, U. (2022, June). Tackling irony detection using ensemble classifiers. In Proceedings of the thirteenth language resources and evaluation conference (pp. 6976-6984).

Uchida, S. (2024). Using early LLMs for corpus linguistics: Examining ChatGPT's potential and limitations. Applied Corpus Linguistics, 4(1), 100089.

Zappavigna, M. (2023). Hack your corpus analysis: how AI can assist corpus linguists deal with messy social media data. Applied Corpus Linguistics, 3(3), 100067.

Logistic regression analysis for function-to-form corpus building

Isabelle Suremann

Zurich University of Applied Sciences

This poster presents a method employing logistic regression analysis to compile a corpus suitable for function-to-form research, with a particular focus on the metadiscursive function, i.e., the function of statements on public discourse that comment on or seek to regulate public discourse and its conditions (Schröter, 2021, 2024; Putterer, 2023). Corpus studies "more commonly begin with a form and, in pragmatic studies, work towards the functional analysis of these forms" (O'Keeffe, 2018: 587). This poses several challenges, e.g. that simple form-based searches are often restricting and do not provide comprehensive results without prior analysis. O'Keeffe (2018) outlines various methods to overcome these issues, including random sampling from a larger corpus to create a manageable subset for functional analysis. Similarly, Bender (2023) suggests starting with a manually annotated corpus sample, which can be scaled to the entire corpus using machine learning algorithms.

Building on these approaches, the proposed method employs logistic regression analysis on a corpus sample to define refined search patterns that can be applied to a larger corpus. First, linguistic features (grammatical and lexical) and the binary variable (metadiscursive: yes/no) are annotated at a paragraph level. Second, logistic regression analyses are conducted to identify significant predictors of the dependent variable. Third, co-occurrence analyses are conducted to identify relevant feature combinations. Fourth, significant positive predictors and relevant combinations thereof are used to identify search patterns in the entire corpus and create a subcorpus. The resulting subcorpus is validated, refined, and eventually used to analyse different categories and types of statements that fulfil the metadiscursive function.

Although labour-intensive, this method has distinct advantages. It helps uncover both linguistic characteristics of the metadiscursive function and is useful in building an adequately specific subcorpus. The poster will cover additional benefits and pitfalls that are currently being explored at the time of writing.

List of references

Bender, M. (2023). Pragmalinguistische Annotation und maschinelles Lernen. In S. Meier-Vieracker, L. Bülow, K. Marx, & R. Mroczynski (Hrsg.), Digitale Pragmatik (S. 267–286). Springer. https://doi.org/10.1007/978-3-662-65373-9_12

O'Keeffe, A. (2018). 23. Corpus-based function-to-form approaches. In A. H. Jucker & W. Bublitz (Hrsg.), Methods in Pragmatics (S. 587–618). De Gruyter Mouton. https://doi.org/10.1515/9783110424928-023

Putterer, E. (2023). Schöne Klimaprosa", „unnötig dramatische Rhetorik" und „Blablabla": Sprachthematisierende Äußerungen und metadiskursive Reflexionen im deutschen Klimawandeldiskurs. Linguistik Online, 123(5), 49–70. https://doi.org/10.13092/lo.123.10549

Schröter, M. (2021). Diskurs als begrenzter Raum. Metadiskurs über den öffentlichen Diskurs in den neurechten Periodika Junge Freiheit und Sezession. In S. Pappert, C. Schlicht, M. Schröter, S. Hermes, C. Riniker, & C. Spieß (Hrsg.), Skandalisieren, stereotypisieren, normalisieren (Bd. 27). Helmut Buske Verlag.

Schröter, M., & Jung, T. (2024). Speaking up and being heard: The changing metadiscourse about 'voice' in British parliamentary debates since 1800. Language & Communication, 94, 41–55. https://doi.org/10.1016/j.langcom.2023.12.002

Automated corpus characterization using LLMs: A register analysis case study

Otto Tarkka, Erik Henriksson, Veronika Laippala, Filip Ginter

University of Turku

Web-scale corpora have become increasingly important in corpus linguistics, but efficient ways to characterize and categorize such massive collections are still lacking. Current approaches require developing task-specific annotation schemes for each research question – whether, for instance, studying register variation, genre evolution, or compiling specific subcorpora. Large Language Models (LLMs) have been shown to be capable of producing human-level annotations in a number of tasks (Gilardi et al., 2023; Yu et al, 2024). Still, despite some early efforts (e.g., Curry et al., 2024; Yu et al, 2024), LLMs are not yet widely utilised in corpus linguistics.

We propose a new methodology for corpus characterization using LLMs to develop comprehensive document descriptors. Using Llama 3.3 70B (AI@Meta, 2024), we allow the model to freely describe documents rather than constraining it to predefined categories. This approach yields rich descriptions covering textual aspects including communicative purpose, situational characteristics, and stylistic features. We test our method on 100,000 documents from FineWeb (Penedo et al., 2024), a web corpus containing diverse online text varieties. Our approach enables both detailed linguistic analysis and creation of task-specific subcorpora without relying on word searches or manual annotation.

As a case study, we examine how our data-driven descriptors relate to registers – text varieties associated with particular situational contexts and co-occurring linguistic features (Biber & Egbert, 2018, Biber & Conrad, 2019). We apply an automatic register classifier (Henriksson et al., 2024) to our FineWeb sample and evaluate how well descriptors predict these categories using both simple statistics (odds ratios, chi-square) and machine learning methods (logistic regression, SVM). Preliminary results show strong predictive performance (micro F1 score up to 0.70). We also investigate how different descriptors appear in texts that blend features from multiple registers, adding to recent work on the fuzzy boundaries between register categories (Biber et al., 2020).

List of references

AI@Meta. (2024). Llama 3.3. https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct.

Biber, D., & Conrad, S. (2019). Register, Genre, and Style. Cambridge: Cambridge University Press.

Biber, D., & Egbert, J. (2018). Register Variation Online. Cambridge: Cambridge University Press.

Biber, D., Egbert, J., & Keller, D. (2020). Reconceptualizing register in a continuous situational space. Corpus Linguistics and Linguistic Theory 16(3). https://doi.org/10.1515/cllt-2018-0086.

Curry, N., Baker, P., & Brookes, G. (2024). Generative AI for corpus approaches to discourse studies: A critical evaluation of ChatGPT. Applied Corpus Linguistics, 4(1). https://doi.org/10.1016/j.acorp.2023.100082.

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. In Proceedings of the National Academy of Sciences, 120(30). https://doi.org/10.1073/pnas.2305016120.

Henriksson, E., Myntti, A., Eskelinen, A., Erten-Johansson, S., Hellström, S., & Laippala, V. (2024). Automatic register identification for the open web using multilingual deep learning. ArXiv preprint. https://doi.org/10.48550/arXiv.2406.19892.

Penedo, G., Kydlíček, H., Ben allal, L., Lozhkov, A., Mitchell, M., Raffel, C., Von Werra, L., & Wolf, T. (2024). The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. In Proceedings of the Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track. https://openreview.net/forum?id=n6SCkn2QaG.

Yu, D., Li, L., Suand, H., & Fuoli, M. (2024). Assessing the potential of LLM-assisted annotation for corpus-based pragmatics and discourse analysis: The case of apology. International Journal of Corpus Linguistics, 29(4). https://doi.org/10.1075/ijcl.23087.yu.

P-frames for pedagogical purposes: Analysing novice writers' output

Benet Vincent[1], **Lee McCallum**[2], **Aysel Şahin K□z□**[3]

[1]Coventry University; [2]University of Edinburgh; [3]Izmir Bak□çay University

This poster reports on ongoing research into the pedagogical utility of phrases retrieved using a phrase-frame methodology in the area of Health Sciences. An earlier study in our project retrieved and analysed p-frames from a corpus of research article introductions in Health Sciences. This enabled us, with the help of experts in this area to compile a list of phrases hypothesised as being useful for novice writers of this genre, in particular those writing in English as a foreign language. The aim of this follow-up study is to ascertain the extent to which the target group for the list is aware of the phrases which have been extracted. We aim to achieve this by collecting work written for publication by inexperienced Health Sciences writers and analysing it using corpus techniques for the presence of phrases on our original list. The poster will discuss this ongoing project and provide details about the size and composition of this new corpus of novice writing. It will also describe the techniques we use to identify what are quite abstract and schematic formulations of phrases (e.g. **[approach/intervention]** ***have/has been*** **(*shown*)** ***to*** **[be effective]**) and it will provide initial findings on these.

A corpus-based study on syntactic fossilization of Wh-Islands

Mengkai Wang

Beijing Foreign Studies University

Wh-islands, some conditions that constrain native speakers from syntactic movement across two same *wh*-feature nodes, have long been studied in syntax (Ross, 1967). Syntactic fossilization indicates a process wherein second language learners fail to fully acquire native-like grammatical structures (Selinker, 1972). This study investigates the phenomenon of syntactic fossilization in the context of *wh*-island structures using a corpus-based approach. Focusing on *wh*-islands, this research aims to uncover patterns of fossilized usage in learner corpora. Firstly, a comprehensive corpus of L2 English learners concerning *wh*-islands and non-*wh*-islands was collected and formed. Secondly, the study identified recurrent errors, avoidance strategies, and deviations in *wh*-island structures by categorizing the structures, inputting the raw data into R environment, and calculating the significance of fossilization. The results were analyzed, and surprisingly syntactic fossilization of *wh*-islands exists in L2 learners robustly, indicating the different syntactic learning mechanisms of native and non-native speakers. The findings highlight the role of linguistic complexity, frequency of input, and first language transfer in shaping fossilization. Additionally, this research examines how proficiency level and learning environment influence the entrenchment of such syntactic patterns. The study provides insights into the cognitive and environmental factors underlying syntactic fossilization, offering implications for SLA theory and pedagogical practices aimed at overcoming fossilized errors in advanced learners.

List of references

Ross, J. R. (1967). Constraints on variables in syntax. Ph.D. Thesis, Cambridge: Massachusetts Institute of Technology.

Selinker, L. (1972). Interlanguage. Product Information International Review of Applied Linguistics in Language Teaching, 10, 209-241.

A corpus-based study of China Men on the deformation of Chinese-American narrative in back translation

Mengkai Wang, Yinghao Lin

Beijing Foreign Studies University

The study conducts a case study of Maxine Hong Kingston's *China Men* to explore the deformation of Chinese narrative in the back translation of Chinese American literature. Chinese narrative, a distinct literary form in Chinese American literature, embodies representations related to China and its culture, such as myths, allusions and cultural symbols. Within the context of increasing global cultural exchanges, the back-translation of Chinese American literature offer unique opportunities to understand cross-cultural dialogues. Previous research have mainly approached Chinese narrative through domestication and foreignization perspectives, showcasing insufficient attention to linguistic aspects, significant dimensions for cross-cultural translation and communications (Gsoels-Lorensen, 2010; Jin, 2013). Western scholars have analyzed Chinese narrative as a narrative technique while Chinese researchers focused on its cultural connotations. In addition, current studies largely center around original text analysis while neglecting the vital role of translation and comparative studies (Li, 2018). The research employs a corpus-based quantitative methodology, based on the BFSU China Chinese-English Parallel Corpus (CECPC) and BLCU Corpus Center (BCC). A self-developed parallel corpus of *China Men* was constructed through procedures, including Label-Studio, EmEditor, CorpusWordParser, TreeTagger, Wordless and CUC_ParaConc, where semi-automatic annotation, text cleaning, alignment and POS tagging are effectively completed. All the data were input into R environment, which accomplished statistical analysis with t-tests or chi-square tests. The results show that the contexts and different cultural backgrounds significantly influence the deformation of back-translation in Chinese-American narrative. This corpus-based study is aimed to provide a deeper understanding of how cultural interaction can function in translation.

List of references

Gsoels-Lorensen, J. (2010). Impossibilized Subjects in Maxine Hong Kingston's China Men: Thoughts on Migrancy and the State of Exception. Mosaic-an Interdisciplinary Critical Journal.

Jin, J. Y. (2013). The Narration of Transnational Territory in Kingston's China Men and Kim's Black Flower. Clcweb-Comparative Literature and Culture.

Li, S. (2018). Cultural Travel and Transformation of "Chinese Themed" English Novel: Analyzing Maxine Hong Kingston's China Men. Foreign Languages and Their Teaching.

A corpus-assisted discourse study of sustainable development in news media: A cross-national comparison of China, the US, and the UK (2012-2023)

Yifan Wang

University of Birmingham

Sustainable Development (SD) is a global imperative, addressing the intricate relationship between environmental well-being and socio-economic issues (Yacoumis, 2018). Understanding SD extends beyond policy analysis to include its media representation. Several studies have examined media discourse on climate change (Boykoff, 2007; Gillings & Dayrell, 2023), but the broader topic of SD remains understudied.

This poster presents preliminary findings from an ongoing larger study employing Corpus-Assisted Discourse Studies (Partington, 2004) to analyze SD media discourse across China, the US, and the UK. The study draws on a corpus of 2,453,077 tokens from news reports published between 2012 and 2023. The current analysis focuses on the Chinese media component, where 100 keywords were extracted through comparison with the English Language Newspapers Corpus (SiBol) and systematically categorized into eight thematic groups.

Preliminary findings from the analysis of the term *sustainable development* within the first category, "core sustainable development terms," reveal distinct patterns in verb usage and agency attribution. Through goal-oriented, promotive, and supportive verbs, Chinese media frames SD as both an achievable goal—requiring active promotion and support—and as a dynamic, ongoing process rather than a fixed state. Furthermore, the analysis highlights consistent patterns in agency attribution, positioning China as a key active agent in SD efforts, particularly in its engagement with developing nations.

This is shown to be in contrast to the same term in the US corpus. In that corpus, the US is not positioned as instrumental in achieving SD, so the corpora are asymmetrical in that respect. In addition, *China* is frequently the focus of negative evaluation in the US corpus, as demonstrated using the Engagement framework (Martin & White 2005).

Future research will: (1) examine the linguistic construction and evolution of SD narratives across these countries, and (2) investigate how media practices and societal contexts shape SD representations.

List of references

Boykoff, M. T. (2007). From convergence to contention: United States mass media representations of anthropogenic climate change science. Transactions of the Institute of British Geographers, 32(4), 477–489. https://doi.org/10.1111/j.1475-5661.2007.00270.x

Gillings, M., & Dayrell, C. (2024). Climate change in the UK press: Examining discourse fluctuation over time. Applied linguistics, 45(1), 111-133.

Martin, J. R., & White, P. R. R. (2005). The language of evaluation: Appraisal in English. Palgrave Macmillan.

Partington, A. (2004). Corpora and discourse, a most congruous beast. Corpora and Discourse, 1, 11–20.

Yacoumis, P. (2018). Making progress? Reproducing hegemony through discourses of "sustainable development" in the Australian news media. Environmental Communication, 12(6), 840-853. https://doi.org/10.1080/17524032.2017.1308405

Tracing concept and lexical patterns in historical language data: A computational approach

Catherine Wong[1], **Susan Fitzmaurice**[1], **Benson Lam**[2]

[1]University of Sheffield; [2]The Hang Seng University of Hong Kong

This research-in-progress investigates the evolution of conceptual and lexical structures in Early Modern English (c. 1400-1800) using the EEBO-TCP corpus and the Linguistic DNA demonstrator. Combining concept modelling (Fitzmaurice, 2021, 2022; Fitzmaurice & Mehl, 2022) and diachronic n-gram analysis, it examines the interplay between discourse-level transformations and formulaic lexical patterns, contributing to a deeper understanding of historical language change (Buerki, 2019, 2020).

As a proof of concept, the study explores three key lemmas: *lord*, *father*, and *king*, illustrating its methodology and preliminary findings. Concept modelling, visualised through Sankey diagrams, reveals thematic networks and lemma traffic within 100-word proximities. For instance, *lord* bridges secular ranks and titles ('earl', 'sir'), personal names ('william', 'thomas'), religious associations ('day', 'sabbath') and familial relations ('daughter', 'marry'). *Father* connects familial terms ('daughter', 'brother') with theological concepts ('holy', 'trinity'), while *king* displays metaphorical links to chess ('pawn', 'bishop', 'check'), reflecting hierarchy and strategy.

In contrast, n-gram analysis highlights diachronic shifts in formulaic expressions. For example, *lord* predominantly associated with religious collocations ('lord jesus christ', 'lord thy god') until the early17th century, with non-religious formulaic patterns only emerging after 1631. *Father* demonstrates stability, maintaining entrenched theological collocations ('father son holy ghost'). Meanwhile, *king* reflects historical and political shifts, with collocates evolving alongside the reigns of specific monarchs. Rather than conveying novel meaning in specific contexts, these MWEs serve as institutional utterances, contextually appropriate within religious doctrine and civic discourse, reinforcing faith and social order.

These examples reveal a gap between conceptual expansions and lexical stabilisations, suggesting distinct patterns of semantic and lexical change. The project demonstrates the potential of computational methods, such as concept modelling and visualisation, to systematically trace these dynamics. It refines diachronic tagging techniques laying groundwork for expanding the dataset to address broader questions about the interplay between conceptual and lexical systems in historical linguistics.

List of references

1. Buerki, A. (2019). Furiously fast: On the speed of change in formulaic language. Yearbook of Phraseology, 10(1), 5-38. https://doi.org/10.1515/phras-2019-0003
2. Buerki, A. (2020). Formulaic Language and Linguistic Change: A Data-Led Approach. Cambridge University Press.
3. Fitzmaurice, S. (2022). From constellations to discursive concepts; or: The historical pragmatic construction of meaning in Early Modern English. Transactions of the Philological Society, 120(3), 489-506.
4. Fitzmaurice, S. (2021). Looking for concepts in Early Modern English: Hypothesis building and the uses of encyclopaedic knowledge and pragmatic work. Journal of Historical Pragmatics, 22(2), 282-300.
5. Fitzmaurice, S., & Mehl, S. (2022). Volatile concepts: Analysing discursive change through underspecification in co-occurrence quads. International Journal of Corpus Linguistics, 27(4), 428-450.
6. Linguistic DNA Project. (n.d.). Linguistic DNA Concept Modelling Demonstrator. Retrieved August 1, 2024, from https://www.linguisticdna.org/cmd/
7. Text Creation Partnership. (n.d.). EEBO-TCP Texts. Retrieved August 1, 2024, from https://github.com/textcreationpartnership/Texts

DR-LIB: CLARIN knowledge centre for digital resources for the languages in Ireland and Britain

Martin Wynne[1], Beatrice Alex[2], Megan Bushnell[1], Mo El-Haj[3], Dawn Knight[4], Will Lamb[2], Mícheál J. Ó Meachair[5], Paul Rayson[3]

[1]University of Oxford; [2]University of Edinburgh; [3]Lancaster University; [4]Cardiff University; [5]Dublin City University

A new CLARIN knowledge centre 'Digital Resources for the Languages in Ireland and Britain' (DR-LIB) was launched in 2024 to provide advice and support to researchers and others who want to find and use software programmes and digital datasets in the languages of Britain and Ireland in all their varieties, in contemporary and historic forms, focussing on native languages, but also other languages as they are used in this region. The 'centre' is a collaboration between CLARIN-UK and researchers in the Republic of Ireland, and is a virtual and distributed network, with a central online presence and contact point with online information to orient and help users, with a growing knowledge base. Queries will be responded to by a network of experts on a best-efforts basis. These experts centred around the CLARIN-UK consortium, plus additional experts in key languages and domains, and experts across Europe in the CLARIN network, and anyone with relevant knowledge and expertise is welcome to join the community and contribute.

The initiative builds on and engages with existing projects, centres of expertise, the Celtic Languages in the Digital Age (CLIDA) initiative, and the UK-Ireland Digital Humanities Association Community Interest Group 'Multilingual DH'. in the initial phases of the operation of the centre, there will be a focus in building knowledge and expertise about digital resources for Celtic Languages and other lesser-resourced languages and varieties. Information about language corpora, lexical data, language models, NLP tools, training sets and other digital language resources are being gathered and added to curated registries and discovery services such as CLARIN resource Families and the Virtual Language Observatory in an ongoing and long-term effort to support the discovery and use of these resources.

Orders of indexicality in online discourse of academic bar: A corpus-based discourse analysis

Zihan Xia, Yifan Li

Department of English and Communication, The Hong Kong Polytechnic University

The emergence of academic bars, which feature academic activities being integrated into bar settings in China's major cities, has sparked intense interest among young people. Online discussions on the phenomenon often delve beyond its surface, revealing complex social and cultural underpinnings. The term "academic bar" has become a multifaceted concept. Drawing on sociolinguistic and linguistic anthropological theories, this study employs the concept of indexicality to examine how the term "academic bar" is associated with social implications. Specifically, it explores the orders of indexicality, which refer to the layered meanings attached to signs (Silverstein, 2003). Following Yoder and Johnstone (2018), this research aims to uncover the various meanings in discourse associated with academic bars and how these meanings are constructed and negotiated by supporters and opponents of academic bars. 160 posts were collected from online responses to the following question on the Chinese Q&A platform Zhihu: "Why are young people enthusiastic about discussing academic topics in bars? What can be gained from participating in academic bars?" After using the word segmentation tool, a corpus containing more than 30,000 words is constructed. By examining the concordance lines and semantic domain clusters, it reveals that the term "bar" is predominantly situated within three indexical orders: bar as a physical entity, a public space for socialising, and a form of social identity with a mixture of intelligentsia and modern lifestyle. Discussions on the third layer are typically embedded with a stance towards the social identity related to academic bars and products of the modern consuming behaviour of the Contemporary Chinese young generation. This study contributes by introducing indexical order in corpus-based discourse analysis and shedding light on the lifestyle and intelligentsia in Contemporary Chinese society.

List of references

Silverstein, M. (2003). Indexical order and the dialectics of sociolinguistic life. Language & Communication, 23(3), 193–229. https://doi.org/10.1016/S0271-5309(03)00013-2

Yoder, M. M., & Johnstone, B. (2018). Unpacking a political icon: 'Bike lanes' and orders of indexicality. Discourse & Communication, 12(2), 192–208. https://doi.org/10.1177/1750481317745753

A systematic literature review of EFL Teacher Talk through the lens of corpus linguistics in the Asian higher education context

Sukmawati Yasim

University of Limerick

This study conducts a systematic literature review (SLR) to investigate Teacher Talk in English as a Foreign Language (EFL) classrooms within Asian universities, using Corpus Linguistics (CL). Employing the PRISMA framework, the review examines empirical studies published between 2014 and 2024. A comprehensive search across multiple databases yielded 856 studies, with 16 ultimately included after rigorous screening and eligibility assessment. The review addresses three key questions: (1) What linguistic patterns characterise Teacher Talk in EFL classrooms in Asia? (2) How does Corpus Linguistics enhance understanding of Teacher Talk's role in student learning? (3) What challenges arise in analysing Teacher Talk using CL tools?. The findings reveal that Teacher Talk in the Asian higher education context is characterised by IRF patterns, scaffolding, questioning, extended learner turns, and wait times. This indicates that six out of 11 features of Teacher Talk Walsh were applied. CL facilitates the empirical analysis of large datasets, providing insights into linguistic patterns and their impact on learning. Challenges include collecting representative classroom data due to diverse educational settings. This study offers valuable insights for researchers, educators, and policymakers aiming to enhance EFL pedagogy in Asia through CL. As Farrell (2020) noted, CL benefits both novice and experienced teachers by equipping them to navigate the complexities of EFL pedagogy.

List of references

Farrell, A. (2020) Corpus Perspectives on the Spoken Models used by EFL Teachers, London & New York: Taylor & Francis.

Walsh, S. (2006) 'Talking the talk of the TESOL classroom', ELT journal, 60(2), 133-141, available: http://dx.doi.org/10.1093/elt/cci100.

Extraction, filtering, and validation of a list of collocations for a learner dictionary of Italian

Fabio Zanda[1], Stefania Spina[1], Irene Fioravanti[1], Luciana Forti[1], Damiano Perri[2], Osvaldo Gervasi[2]

[1]University for Foreigners of Perugia; [2]University of Perugia

This poster illustrates each phase of the selection process of collocation items to be included in a learner dictionary of Italian.

The process involved the extraction of candidate collocations from a ca. 50-million-word reference corpus of Italian encompassing written and spoken texts across 10 registers. A hybrid methodology was employed to combine part of speech (PoS) tagging and syntactic dependency parsing approaches to automatically identify adjacent and non-adjacent collocations (Perri et al., 2024). The grammatical relations extracted were verb-noun, adjective-noun, verb-adjective, verb-adverbial, adverbial-adjective, and noun-noun, producing an initial list of 2 million collocations.

Collocation frequency was calculated individually both for the PoS and parsing methods. When the same collocation was found by both methods, refined frequency counts were computed through a bag-of-words technique (Qader et al., 2019). The collocation list was then integrated by association measures (MI, MI^3, Log-Likelihood, LogDice) and dispersion measures (DP, DP_{norm}) (cf. Gries, 2024).

The filtering process included three steps: Step 1 involved retaining collocations that met minimum thresholds for MI and DP_{norm}. Step 2 focused on identifying collocations using thresholds for frequency and DP_{norm}. Step 3 further filtered the collocations identified in Step 2 by applying a threshold for LogDice values. A manual review of all filtered collocations was conducted to eliminate those with tagging errors, with the final set including 16,820 items.

The validation of the filtered set of collocations was performed by comparing it against two non-corpus-based collocation dictionaries, resulting in a 59.3% of matches in one or both dictionaries. The remaining 40.7% collocations were evaluated for inclusion in the learner dictionary by two groups of human raters, who deemed approximately one third of them suitable.

Overall, 18.7% of the final list of 12,274 collocations were identified and extracted exclusively from the reference corpus, underscoring the pivotal role of corpus methods in lexicography.

List of references

Gries, S. Th. (2024). Frequency, dispersion, association, and keyness: Revising and tupleizing corpus-linguistic measures. Studies in Corpus Linguistics (Vol. 115). Amsterdam: John Benjamins.

Perri, D., Fioravanti, I., Gervasi, O., & Spina, S. (2024). Combining grammatical and relational approaches: A hybrid method for the identification of candidate collocations from corpora. In A. Bhatia, G. Bouma, A. S. Doğruöz, K. Evang, M. Garcia, V. Giouli, L. Han, J. Nivre, & A. Rademaker (Eds.), Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024 (pp. 138–146). Torino, Italia: ELRA and ICCL.

Qader, W. A., Ameen, M. M., & Ahmed, B. I. (2019). An overview of bag of words: Importance, implementation, applications, and challenges. In Proceedings of the 2019 International Engineering Conference (IEC) (pp. 200–204). Erbil, Iraq: IEEE.

Visual-textual interaction in corporate environmental communication: A multimodal corpus analysis of Chinese CSR reports

Shuyue Zhu

University of Southampton

This study will investigate how Chinese companies construct their environmental commitments through multimodal discourse in Corporate Social Responsibility (CSR) reports. Using a novel methodological approach combining Google Cloud Vision API and corpus linguistics tools, I will analyse visual and textual elements in the environmental sections of 168 CSR reports published between 2020 and 2023 by 50 companies across eight energy-intensive industries (including petrochemicals, chemicals, construction materials, iron and steel, non-ferrous metals, paper, electricity, and aviation). A total of 2,678 images were extracted for analysis. Three interrelated corpora will be compiled: (1) a linear text corpus containing environmental discourse, (2) an image annotation corpus with structured data generated via Google Cloud Vision API, and (3) an integrated multimodal corpus that aligns textual and visual content. These corpora will be analysed using *LancsBox X* (Brezina & Platt, 2024), employing tools such as keyword analysis, collocation networks, and concordance searches to uncover patterns of visual-verbal interaction. The research advances the field of business discourse by offering empirical insights into corporate environmental communication in China and demonstrates the analytical potential of integrating AI-powered image recognition with corpus-assisted approaches to multimodal discourse analysis.

List of references

Baker, P. and Collins, L. (2023). Creating and analysing a multimodal corpus of news texts with Google Cloud Vision's automatic image tagger. Applied Corpus Linguistics, 3(1), 100043.

Collins, L.C. and Baker, P. (2024). A computer-assisted analysis of image representations of obesity: comparing UK news content with the World Obesity Federation Image Bank. Visual communication, 0(0) 1–23.

Christiansen A, Dance W and Wild A (2020) Constructing corpora from images and text: An introduction to Visual Constituent Analysis. In: Rüdiger S, Dayter D (eds) Corpus Approaches to Social Media. Amsterdam: John Benjamins, 149–174.

Cultural-linguistic variations in digital health discourse: A computational corpus analysis of self-disclosure and social support dynamics in Chinese and American eating disorder communities

Wenxi Zhu, **Jiayi Chen**

The Hong Kong Polytechnic University

This study employs computational corpus linguistics to investigate the intersection of self-disclosure patterns and social support dynamics within eating disorder (ED) communities on digital platforms, specifically Douban (Chinese) and Reddit (English). Grounded in Communication Privacy Management theory (Petronio, 2002) and social support theory (Feeney & Collins, 2015), the research analyzes 18,769 posts (9,429 Chinese and 9,340 English) and their associated comments using a systematic mixed-methods approach. The methodological framework consists of three phases: (1) the development of a theoretically informed coding scheme for self-disclosure categories (informational, cognitive, affective) and social support dimensions (informational, emotional, esteem, network, tangible); (2) the implementation of fine-tuned multilingual BERT classification, supplemented by inter-rater reliability assessments; and (3) a corpus-driven analysis of linguistic features combining quantitative and qualitative approaches. The findings reveal significant cross-cultural variations in digital discourse patterns. Chinese users predominantly employ indirect-collective narratives characterized by metaphorical expressions and collective pronouns, with a focus on treatment trajectories and professional authority in support-seeking behaviors. In contrast, American users demonstrate emotional-individual storytelling, marked by first-person narratives and explicit body-related vocabulary, favoring peer-based emotional support exchanges. Logistic regression analyses establish robust predictive relationships between these culturally specific discourse patterns and differentiated social support outcomes. These findings extend theoretical understanding of how cultural-linguistic factors shape digital health communication and offer empirical evidence for developing culturally-sensitive online support systems. The study contributes to computational sociolinguistics by demonstrating the efficacy of large-scale corpus analysis in understanding cultural variations in digital health discourse.

List of references

Feeney, B. C., & Collins, N. L. (2015). A new look at social support: A theoretical perspective on thriving through relationships. Personality and social psychology review : an official journal of the Society for Personality and Social Psychology, Inc, 19(2), 113–147. https://doi.org/10.1177/1088868314544222

Petronio, S. (2002). Boundaries of privacy: Dialectics of disclosure. State Univ of New York Pr.

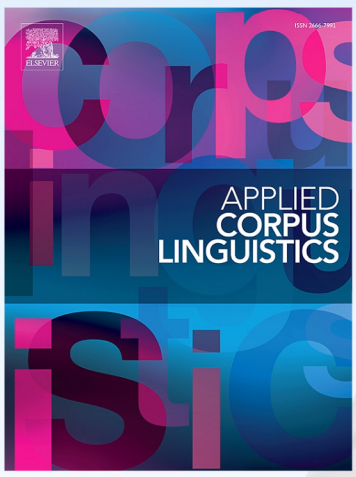
Aston University
BIRMINGHAM UK

BIRMINGHAM CITY University

UNIVERSITY OF BIRMINGHAM

Corpus Linguistics

2025



BLOOMSBURY ACADEMIC

CAMBRIDGE UNIVERSITY PRESS

CLARIN
Common Language Resources and Technology Infrastructure

JOHN BENJAMINS PUBLISHING COMPANY

senso.

SKETCH ENGINE