ICAME

A CONFLUENCE OF CORPUS RESEARCH IN THE AGE OF AI

university of koblenz
Arts / Humanities

KOBLENZ · RHINE & MOSELLE CONFLUENCE

ICAME47 · KOBLENZ · 26–30 MAY 2026

BOOK OF ABSTRACTS

DATES
26–30 May 2026

LOCATION
Koblenz, Germany

HOST
University of Koblenz

wp.uni-koblenz.de/icame47

icame47@uni-koblenz.de

Welcome to ICAME47

Dear colleagues and friends of ICAME,

It is our great pleasure to welcome you to **ICAME47** at the University of Koblenz, 26–30 May 2026.

ICAME (International Computer Archive of Modern and Medieval English) is an annual international conference and one of the longeststanding organisations bringing together linguists and data scientists working with English language corpora. The conference theme is *"A Confluence of Corpus Research in the Age of AI."*

We received a large number of submissions from over 25 countries across all continents, and we are delighted to announce that the programme will feature four plenary talks, five pre-conference workshops, as well as a wide range of full papers, work-in-progress reports, and software demonstrations.

We are honoured to welcome four outstanding plenary speakers: **Laurence Anthony** (Waseda University), **Jonathan Culpeper** (Lancaster University), **Jane Stuart-Smith** (University of Glasgow), and **Natalia Levshina** (Radboud University, the Netherlands). Their talks span the full breadth of where corpus linguistics stands today and where it is heading.

This Book of Abstracts is the scholarly record of ICAME47. We hope it will serve you well during the conference and long after as a reference for the conversations, debates, and discoveries that took place here in Koblenz.

The ICAME47 Organising Committee
University of Koblenz, May 2026


Ann-Katrin Biehl


Nico Boller


Dina Necke


Andreas Weilinghoff


Saran Nair


Jana Semrau


Sarah Wunderlich

Organising Committee

Local Organising Committee

Department of English and American Studies, University of Koblenz

- **Andreas Weilinghoff** (University of Koblenz) — Conference Chair
- **Sarah Wunderlich** (University of Koblenz)
- **Nico Boller** (University of Koblenz)
- **Dina Necke** (University of Koblenz)
- **Ann-Katrin Biehl** (University of Koblenz)
- **Jana Semrau** (University of Koblenz)
- **Saran Nair** (University of Koblenz)
- Student helpers

Scientific Committee

With the support and advice of the ICAME Board and external reviewers

- Laurence Anthony
  Waseda University, Japan
- Gavin Brookes
  Lancaster University, UK
- Rachele De Felice
  Open University, UK
- Stephanie Evert
  FAU Erlangen-Nürnberg, Germany
- Turo Hiltunen
  University of Helsinki, Finland
- Sebastian Hoffmann
  University of Trier, Germany
- Merja Kytö
  Uppsala University, Sweden
- Tove Larsson
  Northern Arizona University, US
- Anna Lindroos Čermáková
  Lancaster University, UK
- Rosa Lorés
  Universidad de Zaragoza, Spain
- Patricia Ronan
  TU Dortmund University, Germany
- Sylvi Rørvik
  University of Inland Norway, Norway
- Edgar W. Schneider
  University of Regensburg, Germany
- Gerold Schneider
  University of Zurich, Switzerland
- Martin Schweinberger
  University of Queensland, Australia
- Jolanta Šinkūnienė
  Vilnius University, Lithuania
- Lukas Sönning
  Universität Bamberg, Germany
- Cristina Suárez-Gómez
  University of the Balearic Islands, Spain
- Benedikt Szmrecsanyi
  Katholieke Universiteit Leuven, Belgium
- Jukka Tyrkkö
  Linnaeus University, Sweden
- Peter Uhrig
  FAU Erlangen-Nürnberg, Germany
- Nuria Yáñez-Bouza
  Universidade de Vigo, Spain

ICAME Board

Executive Board

- Laurence Anthony
  Waseda University
- Gavin Brookes
  Lancaster University
- Rachele De Felice
  Open University
- Stephanie Evert
  FAU Erlangen-Nürnberg
- Turo Hiltunen
  University of Helsinki
- Tove Larsson
  Northern Arizona University
- Anna Lindroos Čermáková
  Lancaster University
- **Patricia Ronan** — *Chair*
  TU Dortmund
- Sylvi Rørvik
  University of Inland Norway
- Gerold Schneider
  University of Zurich
- Martin Schweinberger
  University of Queensland
- Jolanta Šinkūnienė
  Vilnius University
- Cristina Suárez-Gómez
  University of the Balearic Islands
- Benedikt Szmrecsanyi
  Katholieke Universiteit Leuven
- Jukka Tyrkkö
  Linnaeus University
- Nuria Yáñez-Bouza
  Universidade de Vigo

Secretary

Merja Kytö

Uppsala University

Technical Secretary

Peter Uhrig

FAU Erlangen-Nürnberg

The ICAME Journal: 50 Issues

Celebrating 50th Volume of the ICAME Journal

1978 – 2026

The *ICAME Journal* reaches its fiftieth volume in 2026 — a milestone that will be celebrated during a special session on Friday 29 May at ICAME47.

Founded in 1978 as the *ICAME News*, the journal has accompanied the entire arc of modern corpus linguistics: from the first experiments with machine-readable text, through the expansion of corpora in the 1990s, to the computational and AI-assisted approaches that define the field today. As one of the longest-running specialist journals in linguistics, it has served as the primary publication venue for methodological advances, corpus documentation, and theoretical reflections within the ICAME community.

Fifty volumes represent both a record and a responsibility. The *ICAME Journal* has published work that shaped how we design corpora, how we annotate them, how we analyse them, and how we think about what corpora can and cannot tell us.

The celebration takes place on **Friday 29 May 2026, 10:00–10:05** in **E011**.

Practical Information

Venue

University of Koblenz
Room codes: E011 (main auditorium), E113, E114, E313, E314 (parallel session rooms)

Check-in

Check-in desks are open from 8:00am on Tuesday 26 May (workshop day) and Wednesday 27 May.

Social Programme

Conference Warming Wednesday 27 May
Festung Ehrenbreitstein (Koblenz Fortress)
Meeting point: Koblenz Cable Car (City station) at 18:15 h

Boat Trip Thursday 28 May
Meeting point: Deutsches Eck (German corner) at 18:45 h

Conference Dinner Friday 29 May, 20:00 h
Adaccio, Firmungstraße 2, 56068 Koblenz

Abstract Format in This Book

Each abstract is presented in the format submitted by the author(s). The following badges are used throughout:

FULL PAPER Full Paper (20 + 10 min) WIP Work-in-Progress (15 + 5 min) POSTER Poster
SOFTWARE DEMO Software Demonstration WSP Workshop Paper

Contact

Web: wp.uni-koblenz.de/icame47/
Mail: icame47@uni-koblenz.de

Image credits and design notes Photographs of plenary speakers were provided by the respective speakers and are reproduced with their permission. Background and decorative images used in this Book of Abstracts were sourced from Pixabay (https://pixabay.com) and Wikimedia Commons (including the image on the cover — Holger Weinandt, CC BY-SA 3.0 DE, https://creativecommons.org/licenses/by-sa/3.0/de/deed.en). Layout and design were produced using LaTeX with assistance from AI-based design tools.

Contents

RHINE

MOSELLE

Plenary Speakers

Plenary 1 Wednesday 27 May 2026 • 9:30–10:30 • E011

Laurence Anthony

Waseda University, Tokyo, Japan

Laurence Anthony is Professor of Applied Linguistics at the Faculty of Science and Engineering, Waseda University, Japan, and a founding member of the Center for English Language Education in Science and Engineering (CELESE). His research focuses on language data science, AI, corpus linguistics, educational technology, and English for Specific Purposes (ESP). His recent work explores the integration of AI into language research and pedagogy. He is best known for developing the widely used corpus analysis toolkit *AntConc*, for which he received the National Prize of the Japan Association for English Corpus Studies (JAECS).

Advancing Corpus Linguistics with Small, Local, and Multimodal AI Language Models

The foundations of corpus linguistics can be traced back to innovations in computer hardware and software that began in the 1960s. Interestingly, however, the characteristic features of corpus methods have remained largely unchanged. Much of corpus research continues to focus on written text, either directly extracted from written sources or transcribed and annotated from speech. This is despite repeated calls to extend corpus methods to multimodal data. In addition, analysis of corpus data remains largely query-based, relying on formal query languages or regular expression searches. Another consistent feature of the field is an emphasis on transparency, reliability, and replicability in corpus-based research.

In recent years, generative AI (GenAI) has emerged as a major focus of academic and industrial language modeling research. Interestingly, GenAI draws on principles long advocated by corpus linguists such as John Sinclair, but developments in the field have taken a noticeably different path. From the outset, interaction with generative AI systems has been based on natural language rather than formal query syntax. Moreover, while early systems focused primarily on text, there has been a rapid shift toward multimodal processing, with increasing integration of audio, image, and video data. At the same time, AI systems often function as black boxes, producing outputs that can be difficult to evaluate, sometimes unreliable, and often challenging to replicate.

In this paper, I first outline the historical and methodological backgrounds of corpus linguistics and generative AI, highlighting both their shared foundations and key divergences. Next, I examine how the natural language processing capabilities of generative AI models can be integrated into traditional corpus tools to facilitate

searching and support more nuanced analysis and interpretation. Building on this, I then explore how multimodal AI models have begun to bridge the divide between text and other communicative modes and consider how these capabilities can be combined with established corpus techniques to support multimodal corpus analysis. Finally, I discuss the importance of integrating small, locally deployable AI models with corpus tools in order to enhance transparency, improve replicability, and provide effective platforms for teaching and critically evaluating AI-generated outputs.

Throughout the paper, I present concrete examples based on the latest version of the AntConc toolkit, which integrates both large cloud-based and smaller local AI models with traditional corpus methods, offering a unified environment for corpus-based research, teaching, and learning in the age of multimodal AI.

Plenary 2 Thursday 28 May 2026 • 9:00–10:00 • E011

Jonathan Culpeper
Lancaster University, United Kingdom

Jonathan Culpeper is Professor of English Language and Linguistics at Lancaster University, UK. His research spans pragmatics, stylistics and the history of English, all of which he has pursued at some point through corpus methods. A major publication in historical corpus linguistics is *Early Modern English Dialogues: Spoken Interaction as Writing* (2010, CUP; with Merja Kytö). He is currently leading the corpus-based £1 million AHRC-funded Encyclopaedia of Shakespeare's Language project, which will provide evidence-based and contextualised accounts of Shakespeare's language.

Corpus-based explorations in Shakespeare's language and beyond

Compared with the voluminous output of literary critics on Shakespeare's works, linguists have not had much to do with Shakespeare's language. But times are ripe for change. This talk reflects on the corpus-based study of Shakespeare's language. It will shed new light on aspects of Shakespeare's language (and more generally the English language beyond it), and also dispel some myths about Shakespeare's language along the way. I hope that it will additionally be of general interest beyond Shakespeare, not least because of the variety of techniques it deploys and the fact that it engages problematic data (e.g. with irregular spelling) calling for solutions that are also relevant to non-historical data.

I will draw on work undertaken as part of the Encyclopedia of Shakespeare's Language project, a project that deployed a variety of corpus-based methods and resulted in the five-volume *Arden Encyclopedia of Shakespeare's Language* (Culpeper 2023–2026).

Having dispatched the small matter of what exactly is meant by "Shakespeare's language", I first consider the role of Shakespeare's language in shaping the study of the history of English, and especially its effect on the Oxford English Dictionary, illustrating my points with a brief discussion of compound words. I then turn to how Shakespeare's language is viewed by the general public in the context of the English language as a whole. In particular, I consider the myth that Shakespeare created thousands of neologisms.

In the second half of the talk, I examine the linguistic patterns of Shakespeare's language, focussing on affixes, words (and word-meanings) and grammar. I will dwell rather longer on the second of these, as I will also take the opportunity to

introduce the new corpus-based dictionary of Shakespeare's language, Volumes 1 and 2 of the previously mentioned Encyclopedia, and some of the insights it affords. This talk concludes with a coda that addresses the theme of this conference. The Encyclopedia took place over a period of more than 25 years from conception to realisation. I ask myself: if I could go back in time, what would I now have done differently, given the advent of AI?

References

Culpeper, Jonathan (editor-in-chief) (2023–2026, final volume in press). *The Arden Encyclopedia of Shakespeare's Language*. 5 Volumes. London: Bloomsbury.

Plenary 3 Friday 29 May 2026 • 9:00–10:00 • E011

Jane Stuart-Smith
University of Glasgow, United Kingdom

Jane Stuart-Smith has been Professor of Phonetics and Sociolinguistics at the University of Glasgow since 2013, first joining the University in 1997 as Lecturer in English Language, where she has worked with colleagues to develop the Glasgow University Laboratory of Phonetics (GULP). With members of GULP and collaborators outwith Glasgow, she considers the many relationships between speech and society, taking the rich linguistic variation in Scotland as the basis for her work (e.g. *Sounds of the City*). More recently, she has been using corpus phonetic techniques to consider phonetic and phonological variation over space and time in the Englishes of the British Isles and North America (*SPeech Across Dialects of English* — SPADE), and is currently examining Variability in Child Speech in a large longitudinal and cross-sectional cohort of typically-developing children in Scotland.

Are we nearly there yet? Some observations from a corpus (phonetic) traveller across English

Language corpora enable linguists to effectively travel across regional and social varieties of languages, as well as to track language change over time. Previous research provides conceptual maps and motivates research questions to act as routes with which to traverse the actual terrain of our corpora, insofar as the tools at our disposal allow. Until recently, analysis of large amounts of text has substantially outstripped large speech corpus analysis, at least partly because of the complexities of scaling up audio data processing, and achieving robust, consistent, large-scale acoustic phonetic analyses.

In this talk, I discuss how advances in speech data processing combined with large-scale speech corpus analytic software (PolyglotDB, McAuliffe et al 2019), applied to the substantial SPeech Across Dialects (SPADE) meta-corpus (Sonderegger et al 2022; Stuart-Smith et al 2022), help begin to address the question: how well do our maps represent the terrain they are supposed to chart? and specifically: what is 'English' phonology really like? Three case studies drawing on SPADE corpora over space and time, are outlined: the English Voicing Effect (the difference in vowel duration before obstruents signalling the voicing contrast, e.g. beat/bead; Tanner et al 2020), the Scottish Vowel Length Rule (the morphophonological pattern of vowel durations which distinguishes Scottish varieties from all other Englishes; Stuart-Smith and Macdonald 2026), and /r/ in Scottish English (Lawson and Stuart-Smith forthcoming). The findings confirm some assumptions from previous (smaller-scale)

studies, but also show how our maps need to be redrawn. I conclude by taking a detour into different linguistic terrain with a new vehicle, and show how Automatic Speech Recognition, a key tool for automatically processing and segmenting speech corpora, can also act as a novel tool for corpus phonetic analysis.

References

Stuart-Smith, J., Sonderegger, M., Mielke, J., Tanner, J., Willerton, V., & Macdonald, R. (2024, October 8). SPADE – Speech Across Dialects of English. https://doi.org/10.17605/OSF.IO/4JFRM

Stuart-Smith, Jane and Macdonald, Rachel (2026) Scots and Scottish Standard English. In: Hickey, Raymond (ed.) *The New Cambridge History of the English Language: Britain, Ireland and Europe*. Cambridge University Press: Cambridge, pp. 480-509. ISBN 9781009205818 (doi: 10.1017/9781009205818.020)

Lawson, Eleanor and Stuart-Smith, Jane (forthcoming) Rhotics, In Maciej Baranowski, Paul De Dekker and Jennifer Nycz (eds), *Oxford Handbook of Research Methods in Sociophonetics*, OUP: Oxford.

Plenary 4 Saturday 30 May 2026 • 13:30–14:30 • E011

Natalia Levshina
Radboud University, the Netherlands

Natalia Levshina is an assistant professor of communication and computational methods at Radboud University. Her main research interests are linguistic typology, corpora, cognitive and functional linguistics. She also teaches courses on different topics around AI, including Deep Learning and Large Language Models, chatbots and algorithmic bias. She has published *Communicative Efficiency: Language structure and use* (Cambridge University Press, 2022) and is the author of the best-selling statistical manual *How to Do Linguistics with R* (Benjamins, 2015).

Linguistic annotation of corpora in the age of AI

Since the success of ChatGPT and other generative large language models (LLMs), interest in AI within linguistics has grown dramatically, as reflected in the theme of this conference. But what exactly do we mean by AI? In its broad sense, artificial intelligence refers to the ability of machines to perform tasks typically associated with human intelligence, as well as the scientific study of building such systems. Under this definition, corpus linguistics has long relied on AI methods, including tokenisation, part-of-speech tagging, word embeddings and other NLP tools. However, this is not what the term usually denotes today. As captured by Rich and Knight's (1991) classic formulation of the field as "the study of how to make computers do things which, at the moment, people do better", the meaning of AI is a moving target. Drawing on recent corpus evidence from online news and Reddit data, I show that AI has recently shifted its most representative meaning from machine learning algorithms to generative large language and multimodal models, functioning as an autohyponym. Although this semantic flexibility has communicative benefits, it also calls for vigilance against potential misuse (Guest et al., 2025). Moreover, semantic differences are emerging between the short form "AI" and the full form "Artificial Intelligence", following the principles of isomorphism (Goldberg, 1995; Leclercq & Morin, 2023) and communicative efficiency (Levshina, 2022).

Against this backdrop, the talk addresses a central question: how can LLMs be used responsibly in corpus linguistic research? I highlight broader social and environmental implications, including copyright concerns, concentration of power in a small number of technology companies, risks of deskilling, threats to privacy and agency, and the environmental footprint of large-scale computation (Dauner & Socher, 2025). Issues of reproducibility and transparency further complicate the

integration of LLMs into scholarly practice.
These considerations are illustrated through a case study of identification and classification of events in English-language news corpora. Using non-proprietary, locally deployable language models, the study demonstrates how complex annotation tasks (identifying events, distinguishing agentive vs. non-agentive events and identifying explicit vs. implicit agents) can be performed while mitigating social and environmental costs.
By critically examining both the strengths and the limitations of AI in its broad and narrow sense, this plenary aims to contribute to ongoing debates about its role in (corpus) linguistics. I propose a set of desiderata for the responsible use of AI in corpus linguistics, where priority is given to the use of open-weight and local models in combination with well-established NLP tools.

References

Dauner, M., & Socher, G. (2025). Energy costs of communicating with AI. *Frontiers in Communication*, 10, 1572947. doi: 10.3389/fcomm.2025.1572947.

Goldberg, A. (1995). Constructions: A Construction Grammar Approach to Argument Structure. Chicago University Press.

Guest, O., Suarez, M., Müller, B., van Meerkerk, E., Oude Groote Beverborg, A., de Haan, R., Reyes Elizondo, A., et al. (2025). Against the uncritical adoption of 'AI' technologies in academia. Zenodo. https://doi.org/10.5281/zenodo.17065099 .

Leclercq, B., & Morin, C. (2023). No Equivalence: A new principle of no synonymy. Constructions, 15, 1–16.

Levshina, N. (2022). Communicative efficiency: Language structure and use. Cambridge University Press

Rich, E., & Knight, K. (1991). Artificial Intelligence, 2nd edn. McGraw-Hill.

RHINE

MOSELLE

Pre-Conference Workshops

Tuesday 26 May 2026 • University of Koblenz

Workshop schedules were preliminary at time of printing; please check the conference website for any updates.

Workshop : A Confluence of Languages through Corpus-based Contrastive Research in the Age of AI

Room E113 • **9:30–17:30**

Conveners: Signe O. Ebeling (University of Oslo) • Hilde Hasselgård (University of Oslo) • Marie-Pauline Krielke (Saarland University) • Isabell Landwehr (Saarland University) • Sylvi Rørvik (University of Inland Norway)

The workshop takes the general conference theme as its starting point, and combines both the longstanding tradition of having a contrastive workshop at ICAME conferences and the start of a new era for corpus linguistics represented by the emergence of AI. The contributions also bridge the potential gap of traditional corpus linguistics and the influx of AI-inspired topics and methods: Eight of the proposed papers either explore AI-generated datasets or employ AI to supplement traditional corpus methods, while the other papers represent more traditional contrastive corpus-based research. In total, the proposed contributions comprise comparisons of English with eight other languages (Afrikaans, Czech, German, French, Lithuanian, Norwegian, Spanish, and Swedish), and material from a range of registers, including news, promotional discourse, academic prose, and business communication.
Contact: contrastive-icame47@ilos.uio.no

Presentations:

Noun-phrase complexity in English and Norwegian research articles: a comparison of AI-generated and human-authored texts

WSP

Sylvi Rørvik (University of Inland Norway)

This study investigates the use of complex noun phrases (NPs) and noun-phrase modification in the discussion sections of human-authored and AI-generated research articles from the field of education. Previous research has shown that there are differences between human-authored and AI-generated academic prose in several areas, for instance in the accuracy of references (Ariyaratne et al., 2023), in the use of linking adverbials (Briana 2024), in the move structure (Kong & Liu, 2024), in the use of shell nouns (Huang & Deng, 2025), and in lexical diversity and discourse markers, among other things (Pedrawi, 2025). However,

there have not yet been any studies of AI-generated academic prose in Norwegian, so taking a cross-linguistic perspective, the present study focuses on the degree to which there are contrastive differences in the AI-generated discussion sections in English and Norwegian, and the extent to which these match the contrastive differences and similarities in the human-authored discussion sections. A subsidiary aim is to explore whether NP modification is a useful tool to discriminate between human-authored and AI-generated RA discussion sections. The following two research questions form the starting point for the study:

1. To what extent are there differences in in the proportion of complex NPs and in the use of NP modifier types in human-authored and AI-generated RA discussion sections in Norwegian and English?
2. To what extent are there cross-linguistic differences in the proportion of complex NPs and in the use of NP modifier types in human-authored and AI-generated RA discussion sections in Norwegian and English?

The analytical framework is based on Biber et al.'s (2011) overview of phrasal-complexity features. The human-authored material comprises the discussion sections from 10 English and 10 Norwegian empirical research articles from the field of education, which were all were published in peer-reviewed journals in the period 2022-2023. The "corresponding" 10 English and 10 Norwegian AI versions were generated using the large language model (LLM) 'KIchat' (Sikt Norwegian Agency for Shared Services in Education and Research, n.d.), on the basis of prompts containing the RA titles and abstracts.

Preliminary results indicate that the proportion of complex NPs is greater in the AI-generated texts in both languages than in the human-authored ones, but that there are very few other differences between the human-authored and AI-generated texts. There are cross-linguistic differences in the use of several NP modifier types, however, and to a certain extent these occur with the same modifier types in human-authored and AI-generated texts. Further exploration will reveal whether these differences are in the same direction in the AI-generated and human-authored datasets, i.e. whether greater frequencies are found in the same language in both text types.

References

Ariyaratne, S., Iyengar, K. P., Nischal, N., Babu, N. C., & Botchu, R. (2023). A comparison of ChatGPT-generated articles with human-written articles. *Skeletal Radiology*, 52, 1755-1758. https://doi.org/10.1007/s00256-023-04340-5

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development. *TESOL Quarterly*, 45(1), 5–35. https://doi.org/10.5054/tq.2011.244483

Briana, J. C. D. T. (2024). Is ChatGPT-produced text authentic? A contrastive analysis of cohesive markers in human and AI-generated text. Journal of English and Applied Linguistics, 3(2), 38-54. https://doi.org/10.59588/2961-3094.1120

Huang, L., & Deng, J. (2025). "This dissertation intricately explores…": ChatGPT's shell noun use in rephrasing dissertation abstracts. System, 129, 1-16. https://doi.org/10.1016/j.system.2024.103578

Kong, X., & Liu, C. (2024). A comparative genre analysis of AI-generated and scholar-written abstracts for English review articles in international journals. *Journal of English for Academic Purposes*, 71, 1-14. https://doi.org/10.1016/j.jeap.2024.101432

Pedrawi, A. (2025). Linguistic evaluation of content produced by AI and humans in academic texts. TLEP - *International Journal of Multidiscipline*, 2(3), 65-73.

Sikt Norwegian Agency for Shared Services in Education and Research. (n.d.). KI-chat. Available at https://sikt.no/tjenester/sikt-ki/ki-chat. (Accessed September 10, 2025)

Translating the information structure: Indefinite subjects in human and machine translations between English and Czech WSP

Markéta Malá (Charles University in Prague, Czech Republic)

The paper re-visits the questions of constancy of syntactic and information structure in translation between two typologically distinct languages, predominantly analytic English and synthetic Czech. It explores the translation correspondences of the subject, focussing on subjects with non-generic indefinite reference.

Previous studies have shown an overall high degree of interlingual constancy of the subject syntactic function (Dušková 2015, Johansson 2007). Rhematic subjects, however, were found to retain the syntactic function to a lesser extent in both directions of translation between English and Czech (Dušková 2015). This may be accounted for by the different hierarchy of word-order principles in the two languages, with information structure constituting the dominant factor in the Czech relatively free word-order (Firbas 1992), and grammatical principles governing the order of clause elements in English (Biber et al. 2021). In English, the subject tends to be clause-initial, definite and thematic, conveying given/identifiable information. Indefinite subjects in clause-initial position are quite infrequent (Hasselgård 2018), and their post-verbal placement, corresponding to their rhematic character, requires specific syntactic constructions (there-clauses, subject-verb inversion). In Czech, on the other hand, "the subject fairly often assumes the function of the rheme, and stands at the end" (Dušková 2015:15).

Methodologically, the study was inspired by the 'multiple translation project' initiated by Johansson and Øverås (Johansson 2007, Rørvik 2003). In the present research, however, human translators' choices are compared with those made by two machine translation tools frequently used for English-Czech translations, Google Translate and DeepL, addressing the question of whether such translation data could inform cross-linguistic contrastive research (Ebeling 2025), making it possible – like human translations – to explore and highlight recurrent patterns of correspondence between the two languages, and "to differentiate between norm-governed [...] and idiosyncratic behaviour" (Castagnoli 2011:315).

The material was drawn from a fiction sub-corpus of the parallel corpus InterCorp, version 16ud, tagged by language-uniform morpho-syntactic annotation using the universal dependencies framework (Rosen 2023), which makes it easier to compare syntactic structure

across languages.

Two sets of data were analysed. First, English sentences with subjects realised by non-generic indefinite noun phrases and their human translations into Czech were extracted from InterCorp. For each sentence, translations were then created using Google Translate and DeepL. The same steps were followed in the Czech-English direction of translation. The translation counterparts in human and machine translations were compared with respect to the factors whose interplay determines the information load of a clause element: word order, contextual dependence, and semantic structure (Firbas 1992).

The preliminary results confirmed a high degree of constancy of the syntactic function, and similar shifts between the pre-verbal and post-verbal positions of the subject across all translations, human and machine. The machine translations appear to be more alike, and to rely more on the most frequent patterns of correspondence, particularly in clauses with presentative meaning. They may thereby highlight the typical counterparts, but can hardly capture the range of correspondences, as attested in the human translations, which make it possible to explore the role of the subject in the information structure of English and Czech sentences.

References

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (2021). Grammar of Spoken and Written English. John Benjamins.

Castagnoli, S. (2011). Exploring variation and regularities in translation with multiple translation corpora. Rassegna Italiana di Linguistica Applicata, 1-2, 311–332.

DeepL. https://www.deepl.com/en/translator

Dušková, L. (2015). Constancy of the syntactic and FSP function of the subject. In L. Dušková, From Syntax to Text. The Janus Face of Functional Sentence Perspective (pp. 14–29). Karolinum.

Ebeling, S. O. (2015). Corpus-based contrastive studies and AI-generated translations. Languages in Contrast, 25(2), 289–315.

Firbas, J. (1992). Functional Sentence Perspective in Written and Spoken Communication. Cambridge University Press.

Google Translate. https://translate.google.cz

Hasselgård, H. (2018). Sentence-initial indefinite subjects in English and Norwegian, in S. O. Ebeling & H. Hasselgård (Eds.), Corpora et Comparatio Linguarum: Textual and Contextual Perspectives. BeLLS, 9(1), 93–114.

InterCorp, version 16ud (2024). Ústav Českého národního korpusu FF UK, Praha. http://www.korpus.cz

Johansson, S. (2007). Seeing through multilingual corpora: on the use of corpora in contrastive studies. John Benjamins.

Rørvik, S. (2003). Thematic Progression in Translation from English into Norwegian. Nordic Journal of English Studies, 2(2), 245–264.

Rosen, A. (2023). The InterCorp parallel corpus with a uniform annotation for all languages. Jazykovedný časopis, 74(1), 254–265.

Artificial Intelligence or Intelligent Analysis: Using ChatGPT to contrast exclamative constructions

WSP

Lobke Ghesquière & Faye Troughton (University of Mons, Belgium)

The present study intends to explore potential uses of generative AI in corpus-based contrastive linguistics. The burgeoning use of AI in all sectors is of much discussion, none more so than in corpus linguistics, large language models being integral to many of its tools.
Thus far, little research has been conducted into how AI may be used in contrastive linguistics concretely (but note a.o. Ebeling, 2025). Its uses in terms of statistical analysis for any discipline, including linguistics, are clear (e.g. Yu et al., 2024), and its uses in the fields of translation and language teaching are receiving much attention (e.g. Crosthwaite & Baisa, 2023; Lee, 2023). However, studies into the use of AI in the qualitative analysis of corpus data have garnered mixed results. Curry et al. (2024), for example, find that AI can categorise keywords, but performs more poorly in other types of automated qualitative analysis in discourse studies. We intend on pushing this style of investigation further and exploring whether generative AI tools, and more specifically ChatGPT, can categorise larger constructions and compare them across languages.
For this, we have chosen to work with WHAT and HOW exclamative constructions in English, French, and Dutch, as illustrated in (1) to (6). Such examples are easily recognisable as exclamatives, at least to human eyes, but may prove more challenging to AI as their exclamative nature is not always clear from objectively verifiable elements such as punctuation or word order, but may depend on contextualized meaning, polarity and prosody.
(1) What impudence!
(2) How brave you are.
(3) Quelle ambiance !
'What an atmosphere!'
(4) Regardez comme il est fragile.
'Look how fragile he is.'
(5) Wat een kampioen!
'What a champion!'
(6) Hoe onwijs cool is dat?
'How incredibly cool is that?'
(OpenSubtitles 2018)
This study will use ChatGPT-5 to try to replicate the results of a previously conducted contrastive study of English, French and Dutch exclamatives (Ghesquière & Troughton, 2025), and thus to investigate the capabilities of AI at three different tasks. Firstly, it will be asked to retrieve appropriate data from a preselected sample, i.e. to identify exclamatives in three monolingual data sets. Secondly, it will be asked to analyse and contrast the exclamative constructions from all three languages in terms of syntax, polarity, and performative elements. Finally, it will be asked to analyse and contrast quantitative data by performing descriptive and inferential statistical analysis and hypothesis testing. Each task will be

conducted in isolation to ensure that the performance of the AI tool in one task has no impact on any other. It is expected that ChatGPT will handle quantitative data well, but will struggle with identification and contrast. This study may also provide insight into whether its capacities vary from language to language. Given that it was trained on significantly more English data, it may be expected that ChatGPT performs better at handling English data. Having less training data for French and Dutch, lower accuracy in the analysis of data from those two languages may be hypothesized. This may in turn also affect the quality of the results of the contrasting task.

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), 1-9. https://doi.org/10.1016/j.acorp.2023.100082

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

Ebeling, S.O. (2025); Corpus-based contrastive studies and AI-generated translations. Languages in Contrast, 25(2), 289 – 315.

Ghesquière, L. & Troughton, F. (2025) What an echo! A contrastive and parallel study of English, French and Dutch exclamatives using OpenSubtitles 2018 data. Paper presented at ICLC-11, The 11th International Contrastive Linguistics Conference, Prague, 17–19 September 2025.

Lee, S. M. (2023). The effectiveness of machine translation in foreign language education: a systematic review and meta-analysis. Computer Assisted Language Learning, 36(1–2), 103–125. https://doi.org/10.1080/09588221.2021.1901745

Lison, P., Tiedemann, J. (2016). OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC-2016), 2016.

Troughton, F. (2024) Translation and Categorization: The English and French Exclamative. PhD thesis. University of Mons.

Yu, D., Luyang, L., Su, 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), 534 – 561.

Echoes of AI: English, German and Spanish news language before and after ChatGPT

WSP

Isabell Landwehr, Marie-Pauline Krielke, & Sergei Bagdasarov (Saarland University, Germany)

Since their introduction to the wider public, generative AI systems like ChatGPT have changed how we work and study, how we search for information, and how we write texts. In our study, we investigate linguistic contrasts in English, German, and Spanish news texts before and after the release of ChatGPT in November 2022. Using a multi-faceted approach, we detect lexical and grammatical differences in news texts within each language, for the years

2020 to November 2022 (before ChatGPT) and 2023 to 2025 (after ChatGPT). Our main research questions are:
Do we find traceable changes (i.e., typical features of AI-generated language) in news texts after the introduction of ChatGPT?
Are there shared and language-specific patterns?
Previous research, focusing mainly on English, has found certain features to be distinctive of AI-generated texts. For instance, Large Language Models (LLMs) tend to overuse certain words (e.g., delve, emphasize, pivotal) known as excess vocabulary (Kobak et al., 2025; Juzek and Ward, 2025). In general, LLMs (at least older and/or smaller ones) show lower lexical diversity as measured by type-token ratio (TTR) and similar metrics (Culda et al., 2025; Muñoz-Ortiz et al., 2024). When it comes to grammatical features, results vary widely across models and text genres. However, LLMs tend to use a denser and more nominal style and show lower syntactic variability in metrics like sentence and dependency length, tree depth, and branching factor (Bagdasarov and Alves, 2025; Reinhart et al., 2024; Zanotto & Aroyehun, 2025). Studies on German and Spanish are sparse and focus mostly on automatic detection of AI texts without exploring in detail how the distribution of the features differs in human-written and LLM-generated texts (Schaaff et al., 2024; Sarvazyan et al., 2024; Morales-Márquez et al., 2023).
As our dataset, we use the Leipzig Corpora Collection (Goldhahn et al. 2012) encompassing news text corpora for several languages, crawled from freely available daily news portals. We sample from these corpora to create a dataset of 20,000 texts per year and language, from a balanced set of sources and enrich them with syntactic annotations using Stanza (Qi et al. 2020).
For each language, we determine the presence of AI-typical features in both time periods and investigate whether there has been a significant increase. As an exploratory method, we use Kullback-Leibler Divergence (Kullback and Leibler, 1951) to detect salient lexical patterns before and after ChatGPT. We then focus on attested grammatical AI-typical features: TTR, parts of speech, dependency relations, dependency length, and sentence length (Muñoz-Ortiz et al., 2024), and use regression modelling to identify significant differences between pre- and post-ChatGPT texts.
We expect to find significantly more AI-typical features in the texts after November 2022 compared to the previous years. Furthermore, we expect to find converging trends among the different languages. Since the baseline for our features is derived from English data and the vast majority of LLM training data is English (ibid.), we expect to find these features to show up in Spanish and German as well, leading to a stronger divergence from the non-GenAI-influenced texts prior to 2023 since these features are not typical of original German and Spanish texts while English texts should rather show a trend of over-normalization. GenAI-influence may lead to an even stronger divergence for Spanish and German from pre-ChatGPT texts than for English texts.

References

Bagdasarov, S., and Alves, D. 2025. Like a Human? A Linguistic Analysis of Human-written and

Machine-generated Scientific Texts. Proceedings of the First Workshop on NLP and Language Models for Digital Humanities, pp. 38–47.

Culda, L. C., Nerişanu, R. A., Cristesc, M. P., Mara, D. A., Bâra, A., and Oprea, S.-V. 2025. Comparative Linguistic Analysis Framework of Human-written vs. Machine-generated Text. Connection Science (37)1, 2507183.

Juzek, T. S., and Ward, Z. B. 2025. Why Does ChatGPT "Delve" So Much? Exploring the Sources of Lexical Overrepresentation in Large Language Models. Proceedings of the 31st International Conference on Computational Linguistics, pp. 6397–6411.

Kobak, D., González-Márquez, R., Horvát, E. Á., and Lause, J. 2025. Delving into LLM-assisted Writing in Biomedical Publications through Excess Vocabulary. Science Advances 11(27), eadt3813.

Kullback, S., and Leibler, R. A. 1951. On Information and Sufficiency. The Annals of Mathematical Statistics 22(1), pp. 79–86.

Morales-Márquez, L. E., Barrios-González, E., and Pinto-Avendaño, D. E. 2023. Artificial Intelligence-Based Text Classification: Separating Human Writing from Computer Generated Writing. Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023).

Muñoz-Ortiz, A., Gómez-Rodríguez, C., and Vilares, D. 2024. Contrasting Linguistic Patterns in Human and LLM-generated News Text. Artificial Intelligence Review 57, 265.

Reinhart, A., Markey, B., Laudenbach, M., Pantusen, K., Yurko, R., Weinberg, G., and Brown, D.W. 2025. Do LLMs Write like Humans? Variation in Grammatical and Rhetorical Styles. Proceedings of the National Academy of Sciences 122(8), e2422455122.

Sarvazyan, A. M., González, J. A., Rangel, F., Rosso, P., and Franco-Salvador, M. 2024. Overview of IberAuTexTification at IberLEF 2024: Detection and Attribution of Machine-Generated Text on Languages of the Iberian Peninsula. Procesamiento del Lenguaje Natural 73, pp. 421–434.

Schaaff, K., Schlippe, T., and Mindner, L. 2024. Classification of Human- and AI-Generated Texts for Different Languages and Domains. International Journal of Speech Technology 27, pp. 935–956.

Goldhahn, D., Eckart, T., and Quasthoff, U. 2012. Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. Proceedings of the 8th International Language Resources and Evaluation (LREC 2012), pp. 759–765.

Zanotto, S. E., and Aroyehun, S. 2025. Linguistic and Embedding-Based Profiling of Texts Generated by Humans and Large Language Models. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pp. 22852–22869.

Exploring the use of AI-enhanced tools for corpus-based contrastive research: A case study on promotional texts in English and Spanish

WSP

Noelia Ramón, & Belén Labrador (University of León, Spain)

The general availability of electronic corpora in the 1990s represented a breakthrough in linguistic research, including the field of cross-linguistic studies (Aijmer et al., 1996; Johansson & Oksefjell, 1998). Similarly, the advent of generative AI tools is now leading to significant changes in corpus-based contrastive analysis. These changes affect methodological issues,

such as data analysis and the use of data generation tools, and also expand the type of results that can be obtained. Recent studies in this line have compared human and machine translation (Frankenberg-Garcia, 2022), previous corpus-based contrastive results with results from studies based on AI-generated data (Oksefjell Ebeling, 2025), or human corpus annotation with the results of AI-powered annotation tools (Sanz-Valdivieso & López-Arroyo, 2025).

The aim of this paper is to explore the possibilities offered by both traditional corpus-based tools for cross-linguistic analysis and AI-enhanced applications to study promotional discourse in English and Spanish. In particular, we will focus on the differences in the expression of taste in online cheese descriptions. We will extract the empirical data for this study from our English-Spanish comparable corpus of online cheese descriptions (OCD) described in detail in Labrador & Ramón (2024), one of the comparable corpora compiled by the ACTRES research group (actres.unileon.es). The analysis will use a mixed-method approach, incorporating both quantitative and qualitative aspects of how taste is described in English and Spanish. The working procedure will make use of two different types of resources: a) a traditional corpus analysis tool (word lists in the ACTRES browser), and b) several AI-powered tools (same prompt in Notebook LM, Gemini and Perplexity). Abstract key nouns used for the expression of taste (e.g., flavour, tang, finish, in English (Ramón & Labrador, 2018), and e.g., sabor, aroma, notas, in Spanish) will be extracted for the two working languages in each of the four tools. Then, their linguistic context will be explored to identify cross-linguistic similarities and differences. We expect the results will reveal the common denominator in which key nouns are extracted and what the linguistic context is for each noun. The results will also show any additional information provided by each of the tools. The paper will contrast the type of information each tool can provide, as well as differences between English and Spanish.

Preliminary results have shown that traditional corpus tool kits provide quantitative results in corpus analysis using form-based queries as an input, whereas AI applications do not provide this type of information (Labrador, forthcoming). On the other hand, AI tools are designed to answer more open questions, less dependent on form, thus yielding more qualitative findings that complement the linguists' interpretations of the data. The role of researchers in contrastive analysis is changing, and we should now learn to approach research questions differently to make the most of both types of tools and results. The emergence of AI cannot be ignored in corpus linguistics; instead, AI should be embraced as a new source of opportunities that will pave the way for future research in this field.

References

Aijmer, K., Altenberg, B. & Johansson, M. (Eds.). (1996). Languages in contrast. Papers from a symposium on text-based cross-linguistic studies. Lund 4-5 March 1994. Lund University Press.

Frankenberg-Garcia, A. (2022). Can a corpus-driven lexical analysis of human and machine translation unveil discourse features that set them apart? Target, 34(2), 278–308. https://doi.org/10.1075/target.20065.fra

Johansson, S. & Oksefjell, S. (Eds.). (1998). Corpora and cross-linguistic research. Theory, method

and case studies. Rodopi.

Labrador, B. (forthcoming). Capitalizing on genre-based corpora with the use of the AI-powered research tool Notebook LM. Alfinge, Revista de Filología, 37.

Labrador, B. & Ramón, N. (2024). Positive evaluation in the translation of online promotional discourse in the cheese industry. IEEE Transactions on Technical and Professional Communication, 67(3), 316-332. https://doi.org/10.1109/TPC.2024.3417056

Oksefjell Ebeling, S. (2025). Corpus-based contrastive studies and AI-generated translations. Languages in Contrast, 25(2), 289-315. https://doi.org/10.1075/lic.00051.ebe

Ramón, B. & Labrador, B. (2018). Selling cheese online: Key nouns in cheese descriptions. Terminology, 24(2), 210-235. https://doi.org/10.1075/term.00019.ram

Sanz-Valdivieso, L. & López-Arroyo, B. (2025). Human vs. ChatGPT corpus annotation: Data augmentation using LLM fine-tuning. In R. Rabadán & N. Ramón (Eds.), Cross-linguistic mediated communication. Hybrid text production English-Spanish. Peter Lang. (pp. 111-133).

Moves, Stance and Engagement in the Concluding Sections of Research Articles in Lithuanian and English in Two Disciplines WSP

Jolanta Šinkūnienė (Vilnius University, Lithuania)

Ever since John Swales (1990, 2004) introduced and refined the concept of "moves" and "steps" in research article introductions, there has been a growing body of research into how scholars from different disciplines and different cultures structure their academic texts (for an overview see Casal & Kessler (2023)). Much of the existing research, however, has focused on the introductory section of the research article and on texts written in English. Even though the importance of the introductory section in research articles is undeniable, the relevance of the concluding section is equally substantial, as it provides a conceptual summary of the results, gives a high-impact message to the reader and maps the directions for future research. It is rather surprising, therefore, that little research so far has focused on the concluding section of a research article, a few exceptions being Moritz et al. (2008), Adel & Ghorbani Moghadam (2015), Zamani & Ebadi (2016), Farneste (2017), Alkamillah et al. (2022). Moreover, apart from carefully crafted rhetorical moves, the concluding section also calls for a well-considered balance of stance and engagement markers, helping the authors to sound reasonably convincing and engaging in providing the final overview and implications of their results. Research on these markers in languages other than English is quite limited, despite the fact that interesting cross-linguistic differences have been discovered (see, for example, Loi et al. (2016)).

The aim of this paper is to identify and compare rhetorical moves in concluding sections of research articles written in two languages (Lithuanian and English) and in two core disciplines representing humanities and social sciences (linguistics and sociology), as well as to determine which markers of author stance and reader engagement are characteristic to these moves. The study is based on a self-compiled corpus of 40 concluding sections

(10 from each discipline and language) and applies qualitative and quantitative analysis using Moritz et al.'s (2008) detailed rhetorical moves framework and Hyland's (2005) Stance and engagement model. Since the employment of stance and engagement markers is largely determined by the cultural background of the authors, only articles written by native speakers of Lithuanian and English were selected for the study. Following Sala (2008), several criteria were used to verify the cultural background of the authors. The author had to be affiliated to an institution in a country where English is the national language. Also, the bio information of the authors was verified through web searches, and if the author was educated in an English speaking country, they were considered native speakers of English. The same procedure was applied for Lithuanian authors.

The preliminary results show that the rhetorical structure of moves in the concluding sections of the analysed articles is more elaborate in the English language texts irrespective of discipline. The English language conclusions are also much more promotional, employing more and more varied stance and engagement expressions than the analysed texts in the Lithuanian language. This suggests that the intended audience and the context of the publication (international vs local) may play an important role in rhetorical strategies employed by scientific writers. The study may have important implications for teachers and students of English for academic purposes, as well as for non-native English researchers seeking to publish in English in international journals.

References

Adel, S. M. R., & Ghorbani Moghadam, R. (2015). A comparison of moves in conclusion sections of research articles in psychology, Persian literature and applied linguistics. Teaching English Language, 9(2), 167–191.

Alkamillah, M., Azwandi, A., & Maisarah, I. (2022). The conclusion sections in applied linguistics international journal articles written by Indonesian authors. Journal of Applied Studies in Language, 6(2), 118–130.

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.

Farneste, M. (2017). Moves in the sections "Conclusion" and "Conclusions" in applied linguistics research articles. Baltic Journal of English Language, Literature and Culture, 7, 58–73.

Hyland, K. (2005). Stance and engagement: A model of interaction in academic discourse. Discourse studies, 7(2),173–192.

Loi, Ch.-K., Lim, M-H. K., Wharton, S. (2016). Expressing an evaluative stance in English and Malay research article conclusions: International publications versus local publications. Journal of English for Academic Purposes, 21, 1–16.

Moritz, M. E. W., Meurer, J. L. & Kuerten Dellagnelo, A. (2008). Conclusions as components of research articles across Portuguese as a native language, English as a native language and English as a foreign language: A contrastive genre study. The ESPecialist, 29(2), 233–253.

Sala, M. (2008). Argumentative styles as cultural identity traits in legal studies. Linguistica e Filologia, 27, 93–113.

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.

Zamani, G., & Ebadi, S. (2016). Move analysis of the conclusion sections of research papers in Persian and English. Cypriot Journal of Educational Sciences, 11(1), 9–20.

Making sense of GenAI for contrastive purposes: Verbs of sensing in English and Norwegian

WSP

Signe Oksefjell Ebeling, & Jarle Ebeling (University of Oslo, Norway)

As part of a larger study of heterosemous, cognate nouns and verbs in English and Norwegian (Ebeling & Ebeling, In prep.), the verb pair sense and sanse was found to differ considerably in frequency of use: sense is six times as frequent as sanse in modern fiction. This observation prompted us to look more closely at verbs of sensing in the two languages, since Norwegian must have verbs other than sanse that correspond to the meanings and uses of sense. The bidirectional data used in the study referred to above uncovered the following four Norwegian verbs as relatively frequent correspondences of sense: ane ('perceive'), fornemme ('be aware of'), føle ('feel') and merke ('notice'). These Norwegian near synonyms of sanse and their English counterparts are in some way or other related to the core meaning of sense and sanse, i.e. "to perceive (something) by the senses; become aware of" (Dictionary.com). However, it is not clear exactly how the verbs interconnect, i.e. under what conditions do they overlap in terms of lexico-grammar and under what conditions do they not.

To disentangle the differences in meaning and use of these near synonyms within and between the two languages we will test the potential of GenAI for contrastive purposes, seeking an answer to the following research question:

How do the meaning and behaviour of verbs of sensing in English and Norwegian fiction relate to each other within and across the two languages and to what extent can GenAI help us in the process?

The material for the study is taken from two monolingual corpora: the Corpus of British Fiction and Leksikografisk bokmålskorpus. To work with as comparable data as possible we restricted the material to approx. 13 million words of modern fiction (2000-2012) from each corpus. The in-depth contrastive analysis is based on 100 randomly extracted concordance lines of each verb. These lines were fed into two Chatbots for analysis (Copilot Chat and Chat-GPT-5) and were also scrutinised and analysed manually.

Preliminary findings suggest that, at a general level, the Chatbots produce relatively sound insights regarding the lexico-grammatical characteristics of these verbs as well as the interrelation between them, both cross-linguistically and within each language. However, and in line with previous observations of GenAI corpus analysis (e.g. Curry et al. 2024), despite being given authentic corpus material, the examples are often misrepresented, notably by changing the tense (e.g. from present to past), the verb (e.g. from sanse to ane), and by presenting each example as a full sentence. Even with access to a specific number of

examples, neither Chatbot gives frequency information unless prompted to do so, and the counts tend to be imprecise. These observations indicate that the underlying LLMs may be more important for the Chatbots than corpus material provided with the prompt when producing a linguistic analysis, contrastive or otherwise.
Once we have carried out the full analysis of the verbs, we will also provide a more detailed account of the contrastive insights gained from this study.

References

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

Ebeling, J. & Ebeling, S.O. (In prep). A contrastive study of heterosemous cognate noun/verb pairs in English and Norwegian.

"Sense" (verb), Dictionary.com (2025). https://www.dictionary.com/browse/sense.

Attitudinal stance adverbials in English and Norwegian

WSP

Hilde Hasselgård (University of Oslo, Norway)

This paper investigates stance adverbials in English and Norwegian. A subsidiary aim is to try out ways in which Gen-AI tools might assist in various steps of the research procedure and to compare its performance to traditional (human) analysis.
Stance adverbials, or disjuncts, convey a speaker's or writer's comment on some aspect of the content of a message or the style in which it is presented (Biber et al. 1999: 764; Quirk et al. 1985: 615). Disregarding stance adverbials denoting mainly epistemic modality (e.g. *probably, maybe*) or relating to the form of the message (e.g. *briefly / kort sagt*), this study considers adverbials conveying the speaker's attitude to or judgment of the content of a proposition, as in (1) and (2). Such *attitude adverbials* (Biber et al. 1999) or *comment adjuncts* (Halliday & Matthiessen 2014) can be content-, event- or participant-oriented (Kinn 2023).

(1) *Ironically*, this reduces file size, too. <ICE-USA:W2B>
(2) Det er *utrolig nok* sant! <ICC-NO-E1B-007> ["It is incredibly enough true."]

The material comes from two comparable corpora of written English and Norwegian: the US component of the International Corpus of English (ICE-US) and the Norwegian component of the International Comparable Corpus (ICC-NO). The comparable corpus approach precludes the identification of corresponding terms through translation relations; hence the sets of search terms were based on lists of stance adverbials from grammars and previous studies and on translations of these lists provided by ChatGPT.
The research questions are the following: What types of attitude adverbials are found in the two corpora? Are the types, their frequencies, and their usage patterns the same in English and Norwegian? A subsidiary methodological question is what parts of a contrastive corpus

analysis (based on comparable corpora) Gen-AI might assist with.
Attitudinal stance adverbials have been studied much less than the epistemic ones. They are less frequent than epistemic adverbials (Biber et al. 1999: 859) and can be hard to classify (Halliday & Matthiessen 2014: 190; Kinn 2023). Furthermore, Hasselgård (2012), in a study based on bidirectional translation data, found that categories of stance adverbials sometimes overlap in a cross-linguistic and translation perspective.
In the analysis, ChatGPT5 is tried out as a potential aid in the classification of attitudinal adverbials in isolation and as part of concordance lines (similarly to Curry et al. 2024, see also Anthony 2025). The prompt is accompanied by a reduced concordance for each language with only one token per adverbial type, since an early attempt indicated that ChatGPT does not cope well with long concordances where the keywords vary greatly in frequency. First, ChatGPT was prompted to suggest functional categories for the keywords and second, to take account of the left and right context in the classification. The categories suggested by ChatGPT were similar, but not identical, for the two languages, and appeared more fine-grained for English. However, ChatGPT does not seem to take account of context from concordance lines even when instructed to do so, basing the classification on superficial rather than contextual meaning.

References

Anthony, L. (2025). Concordancing with AI: Applications of word and sentence embeddings. *Applied Corpus Linguistics* 5(3).

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). *Longman grammar of spoken and written English*. Longman.

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.

Halliday, M.A.K and Matthiessen, C.M.I.M. (2014). *Halliday's introduction to functional grammar*. Routledge.

Hasselgård, H. (2012). Kommentaradjunkter i kontrastivt perspektiv. En korpusbasert studie. In Andersen, T. & Boeriis, M. (eds), *Nordisk Socialsemiotik. Pædagogiske, multimodale og sprogvidenskabelige landvindinger*, 177–198. Syddansk Universitetsforlag.

Kinn, T. (2023). Interessant nok: ein produktiv setningsadverbial-konstruksjon. *Norsk Lingvistisk Tidsskrift* 41, 119–145.

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. (1985). *A comprehensive grammar of the English language*. Longman.

Corpora and Open AI tool

International Comparable Corpus, Norwegian component (ICC-NO). See `https://www.hf.uio.no/ilos/english/services/knowledge-resources/icc-no/`

International Corpus of English, USA (ICE-US). See `https://www.ice-corpora.uzh.ch/en.html`

ChatGPT5, accessed from `https://gpt.uio.no/`.

And the Reason Is…: A Cross-Linguistic Investigation of The English Patterns 'The reason is because / The reason is that' and the corresponding Norwegian patterns 'Grunnen er fordi / Grunnen er at.'

WSP

Øyvind Thormodsæter, & Signe Laake (Oslo Metropolitan University, Norway)

The pattern illustrated in examples such as "The reason I'm calling is because I need your help with something" is described as ungrammatical in most prescriptive accounts (see for example Cherry (1933, p. 56,) Blamires (2000, p. 207)). Arguably, the pattern is tautological, and represents a grammatical mismatch in that an adverbial clause takes on a nominal function. However, Hirose (1991, p. 1) notes that "Despite prescriptive objections to using because for that, this use is widely seen both in speech and in writing, and it is also mentioned in one way or another in most, if not all, English usage guides and dictionaries". Furthermore, Bolinger (2022, p. 198) points out that similar tautologies are found in patterns like 'The time […] was when' or 'The place […] is where' without causing similar objection. Several LLMs, including ChatGPT and CoPilot, suggest that 'the reason BE' is considered informal and unacceptable in written contexts, but is used widely in informal contexts.

Although the debate has perhaps not been as outspoken in Norway, Hegge (2015) notes in his newspaper column the tautological nature of 'Grunnen er fordi'. Meanwhile, Faarlund et al. (1997) acknowledge similar patterns in Norwegian, but does not mention 'Grunnen er fordi' specifically.

Investigating a hypothesis that the patterns mentioned above are becoming more frequently used in English and Norwegian, respectively, the current presentation reports the following research questions:

How frequent are the patterns REASON BE because/that and GRUNN VÆRE fordi/at in English and Norwegian, respectively?

Are there indications of changes in the frequency of use over time?

Are there any recurring lexico-grammatical features associated with each pattern?

The English data are retrieved from the BNC 1994 and the BNC 2014, while the Norwegian data are extracted primarily from Leksikografisk Bokmålskorpus 'The Lexicographic Corpus of Bokmål Norwegian' (LBK). LBK has 50 % translations, and the texts are from partly different decades and genres compared to the BNCs, so it is not directly comparable with either of the English corpora. Nevertheless, we include some cross-linguistic observations about frequency of occurrence across the three corpora, albeit without drawing any strong conclusions. Additionally, we will present findings from data compiled from more comparable parts of BNC 1994 and LBK, including an analysis of the lexico-grammatical features of the word combinations occurring on either side of the copula verb.

The findings suggest that REASON BE because is becoming more frequent in English, and that the pattern is more frequent in spoken contexts than in written ones. Furthermore, both REASON BE because and GRUNN VÆRE fordi are more frequent with 3-7 words occurring in between REASON and BE. Finally, GRUNN VÆRE fordi is much less frequent in the Norwegian data than REASON BE because is in the English data, and we assume that

this is at least partly because the LBK data is exclusively written, and not as recent as the BNC2014 material.

References

Blamires, H. (2000). The Penguin Guide to Plain English: Express Yourself Clearly and Effectively. Penguin Books.

Bolinger, D. (2022). Language – The Loaded Weapon: The Use and Abuse of Language Today. Classical Edition. Routledge.

Faarlund, J. T. Lie, S and Vannebo, K.I. (1997) Norsk referansegrammatikk. Universitetsforlaget, Oslo

Hegge, P. I (2015,15 august). Grunnen fordi. Aftenposten, https://www.aftenposten.no/meninger/i/ap37/grunnen-fordi

Hirose, Y. (1991). "On a certain nominal use of because-clauses: Just because because-clauses can substitute for that-clauses does not mean that this is always possible". English Linguistics 8, pp. 16-33. Tokyo: The English Linguistic Society of Japan.

Corpora:

British National Corpus, version 3 (BNC XML Edition). 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/; CQP-edition version 4.0; The CQP-edition of BNCweb was developed by Sebastian Hoffmann and Stefan Evert, accessed via http://www.tekstlab.uio.no/bnc/BNCquery.pl?theQuery=search&urlTest=yes.

Leksikografisk Bokmålskorpus 'Lexicographical Corpus of Bokmål Norwegian'. Knudsen, Rune Lain & Fjeld, Ruth Vatvedt: LBK2013: A balanced; annotated national corpus for Norwegian Bokmål. Proceedings of the workshop on lexical semantic resources for NLP at NODALIDA 2013; May 22–24; 2013; Oslo; Norway. NEALT Proceedings Series 19. Accessed through the Glossa interface at: https://tekstlab.uio.no/glossa2/bokmal

South African English and Afrikaans past time reference: contrasts and competition in contact since the 19th century

WSP

Ronel Wasserman, Bertus van Rooy, & Megan van Winsen (University of Amsterdam, The Netherlands)

In this paper we compare the use of the perfect aspect and simple past for reference to situations from the past in 19th and 20th century corpora of South African English and Afrikaans (still regarded as Cape Dutch in the 19th century). Present-day Afrikaans grammar has no inflected past tense or preterite anymore (although some inflected forms are preserved in a handful of verbs), and has instead repurposed the analytical present perfect form that existed alongside the simple past in the 17th and 18th century Dutch input variety to be the principal way of referring to situations in the past, in the construction het+ge-V (Raidt, 1991). This change occurred mainly during the 18th and 19th century and was completed in the first half of the 20th century (Deumert 2004; Kirsten, 2013; Raidt, 1991).

This overlaps largely with the period during which Afrikaans increasingly came into contact with English, often under complex circumstances. Whereas many other linguistic changes during the two centuries of contact have led to convergence between the two languages (e.g. Wasserman 2016, 2019), this change appears to have increased the distance between them. We wish to not only contribute to the description of the grammatical development of Afrikaans, but also to understand possible reasons for the unexpected divergence between English and Afrikaans.

The main data for Afrikaans comprises five sets of corpora, namely the a collection of mostly business correspondence from the Cape Archives spanning the 18th and early 19th century (Scholtz, 1980), the Corpus of Cape Dutch Correspondence, which contains mostly private letters written in the Cape from the 1880s to 1920s (Deumert 2004), the Nienaber Collection of news and magazine articles from the Cape, Free-State and Natal from the 1830s to 1860s (Nienaber 1942), a corpus of personal letters written between 1899 and 1902 during the Anglo-Boer War (Nel 2016), and a corpus of fiction, non-fiction, academic texts, and manuscripts spanning the 1910s to 1980s (with some data for the 21st century) (Kirsten 2019). The Historical Corpus of South African English (Wasserman 2019) serves as the comparative dataset for South African English from the 1820s to 1950s, and consists of private and business letters, news reportage, fiction and non-fiction from a range of locations.

Some additional data from fiction and news reportage for both Afrikaans and South African English are included to represent the mid- to late 20th century, which was the period when contact between the two languages peaked in the midst of social segregation. The same sampling procedure adopted by Wasserman (2019) and Kirsten (2019) is adopted for the new data. Our analysis compares the development of past time reference over time for both languages, aided by the automated part-of-speech tagging for the English data and manual tagging for the Afrikaans data. A comparison of forms (synthetic and analytic) is supplemented by semantic analysis of random samples of all formal variants to determine the temporal and aspectual properties of the past-time references. Random samples of 200 forms per half-century are be analysed for each of the two languages. Typical semantic characteristics of the English (present) perfect aspect and the Dutch voltooid tegenwoordige tijd (completed present tense) will be used to establish degree of correspondence with the prototype present perfect uses of the colonial sources, such as current relevance, resultative readings, and incompleteness at reference time. The typical semantics of simple past tense reference in English and onvoltooid verleden tijd (incomplete past tense) will serve as starting point for the analysis of the synthetic forms, such as the location of an event at a past moment in time, as well as hypothetical and stance readings. However, both sets of semantic options, alongside possible innovative uses in the data will be considered as the potential semantic space for all forms in the sample. While we examine the historical data to compare the divergent development of past reference in the two languages, we also assess the possibility of convergence between the semantic scope of the English present perfect and the expanded Afrikaans analytical past tense form.

References

Deumert, A. (2004). Language Standardization and Language Change: The Dynamics of Cape Dutch. Amsterdam: John Benjamins.

Kirsten, J. (2019). Written Afrikaans since Standardization: A Century of Change. London: Bloomsbury.

Nel, A. (2016). 'Ons is nog alen Vres door Den Zegen Des Heeren': Gesondheidsformules in Afrikaans-Nederlandse persoonlike briewe uit die Anglo-Boereoorlog. MA dissertation. Vanderbijpark: North-West University.

Nienaber, G.S. (1942). Afrikaans tot 1860. Patriotvereniging vir Afrikaanse Teksuitgawes, nr. 6. Johannesburg: Voortrekkerpers Beperk.

Raidt, E.H. 1991. Afrikaans en sy Europese Verlede. Derde uitgawe. Kaapstad: Nasou.

Scholtz, J. du P. 1980. Wording en ontwikkeling van Afrikaans. Kaapstad: Tafelberg.

Wasserman, R. 2016. Moet en must: 'n geval van Afrikaanse invloed op Suid-Afrikaanse Engels. Tydskrif vir Geesteswetenskappe, 56(1): 25-44.

Wasserman, R. 2019. Historical development of South African English: semantic features. In: Hickey, R., ed. English in Multilingual South Africa. Cambridge: Cambridge University Press. pp. 52-73.

Workshop : Corpus-Linguistic Approaches to Reddit Room E313 • 9:00–15:30

Conveners: Daria Dayter (University of Tampere) • Thomas C. Messerli (University of Basel) • Sven Leuckert (Technische Universität Dresden)

Reddit is a prominent social platform on which millions of international users gather every day to discuss an enormous range of topics in over a hundred thousand different communities called 'subreddits'. The goals of the workshop are (1) to showcase the potential of language use on Reddit specifically for corpus-linguistic approaches; (2) to highlight the multifaceted nature of communicative practices on the platform; and (3) to open up avenues for further collaboration. Contributors to the workshop present case studies of their research on Reddit, emphasising methodological challenges and questions. These include, among others, increasingly AI-produced contributions on Reddit, changes in how data may be accessed and reproduced, and research ethics. Overall, the workshop serves to bring together corpus linguists working with Reddit data and to advance social media research in linguistics. The panel features eight talks that exemplify the diversity of corpus-linguistic approaches to Reddit discourse. Together, these contributions discuss the richness of Reddit as a site for corpus-based linguistic inquiry.
Contacts: daria.dayter@tuni.fi, thomas.messerli@unibas.ch, sven.leuckert@tu-dresden.de

Presentations:

---

thatsthejoke.jpg – Affordances and referentiality on Reddit WSP

Max Hoferichter (University of Greifswald, Germany)

Making references is an important and significant practice on many social media sites, but especially so on Reddit. As early as 2014, the forum website was described to have "transformed itself from a dedicated gateway to the Web to an increasingly self-referential community" (Singer et al. 2014: 1). Accordingly, there is an abundance of meta-level posts and memes that, in turn, refer to these referential practices: a fitting example is the widely-used "I understood that reference" reaction image depicting superhero Captain America (`https://knowyourmeme.com/memes/i-understood-that-reference`), which itself has become the object of a multitude of references (see, e.g., `https://www.reddit.com/r/Showerthoughts/comments/1g15c1v/i_understood_that_reference_is_a_reference_that/`). In this web of allusions, a variety of recurring practices have emerged, one of which sees users refer to this and other memes in text comments by the use of ostensible file names such as "I understood that reference.jpg" (e.g., `https://www.reddit.com/r/Jokes/comments/1m80vcq/comment/n4xeu9a/`). Thus, technological materialities are transposed to new contexts and are employed as referential shortcuts. They thereby become exemplary instances of "social media affordances [...], emerging through the relation of technological, social, and

contextual" (Ronzhyn, Cardenal & Batlle Rubio 2022: 3178). This interplay – between the oftentimes rigid structures of websites and the dynamic navigation of those spaces by their users – is the focus of this report: it will examine how Reddit users recontextualize material from a variety of technological sources in order to reference items of shared knowledge, to express themselves, and to form groups. The analysis will be based on a corpus of pertinent Reddit posts and comments, whose creation which will also be discussed from a methodological perspective. This is due to the challenges that arise when examining artifacts defined by fixed and mutable parts and deployed in a space where "variation often embeds social meaning" (Zhou/Jurgens/Bamman 2024). Firstly, the vast variability observed in the linguistic data necessitates a qualitative approach complement the quantitative data. That way, overarching patterns among the varied examples can be solidified into a typology of practices. Secondly, different strategies of data collection have to be employed: due to the extreme depth some of the referential rhizomes reach, those especially oblique acts of reference are best searched for not directly, but by the metapragmatic responses (of understanding, bewilderment or the need for clarification) they elicit. Thirdly, the difficulties described here show the importance of an approach informed not only by corpus methods, but also by insights into the community and their unwritten rules gained via the methods of digital ethnography (Varis 2015).

References

Ronzhyn, A., Cardenal, A. S., & Batlle Rubio, A. (2022). Defining affordances in social media research: A literature review. New Media & Society, 25(11), 3165–3188. https://doi.org/10.1177/14614448221135187

Singer, P., Flöck, F., Meinhart, C., Zeitfogel, E., & Strohmaier, M. (2014). Evolution of Reddit: From the Front Page of the Internet to a Self-referential Community? arXiv. https://arxiv.org/abs/1402.1386

Varis, P. (2015). Digital ethnography. In A. Georgakopoulou & T. Spilioti (Eds.), The Routledge Handbook of Language and Digital Communication (pp. 55–68). Routledge. https://doi.org/10.4324/9781315694344

Zhou, N., Jurgens, D., & Bamman, D. (2024). Social Meme-ing: Measuring Linguistic Variation in Memes. Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 3005–3024. https://doi.org/10.18653/v1/2024.naacl-long.166

dOnT uSe ThAt WeIrD sPoNgEbOb MeMe: Alternating caps as a strategy for voice management in digital discourse

WSP

Axel Bohmann (University of Cologne, Germany)

Alternating caps refers to a creative spelling strategy in which uppercase and lowercase characters are used in alternation, irrespective of standard orthographic conventions. While its use is attested since the early days of the internet, the strategy was popularized by a

meme in 2017 (Hathaway 2017). At this time, its functional meaning – to express sarcasm or ridicule – also became conventionalized. In its present form, alternating caps is thus a strategy for the lamination of voices, where a statement is presented while at the same time the author indicates a critical distance to the propositional content of the statement. Creative spelling in this context serves as a contextualization cue (Gumperz 1982) that would, in spoken interaction, be provided by pitch movement, facial expressions, and other channels unavailable in text-based digital discourse (cf. Klewitz & Couper-Kuhlen 1999). This talk traces the development of alternating caps on the web forum Reddit (cf. Messerli et al. 2025), considering its rise in frequency, its spread across contexts (subreddits, topics), and developments in its formal nature (reliance on visual support and on direct quotes to echo). Over time, it is shown that alternating caps can be used with increasing flexibility, requiring neither reference to the meme template the practice is derived from nor to any specific, previously produced utterance to be rephrased. The process of spread is akin to the indexical bleaching Squires (2014) observes for a specific quote by a TV personality. The talk discusses the functions of alternating caps in the wider context of quotation and voice management online and in relation to reported speech and voice-in-contrast in offline sociolinguistic work (Agha 2005; Bucholtz 1999).

References

Agha, A. (2005). Voice, footing, enregisterment. Journal of Linguistic Anthropology, 15(1), 38–59.

Bucholtz, M. (1999). You da man: Narrating the racial other in the production of white masculinity. Journal of Sociolinguistics, 3(4), 443–460.

Gumperz, J. J. (1982). Discourse Strategies. Cambridge: Cambridge University Press.

Hathaway, J. (2017). "Mocking Spongebob" is the most insulting meme of 2017. The Daily Dot. https://www.dailydot.com/memes/mocking-spongebob-meme/.

Klewitz, G. & Couper-Kuhlen, E. (1999). Quote – unquote? The role of prosody in the contextualization of reported speech sequences. Pragmatics, 9(4), 459–485.

Messerli, T., Dayter, D., Leuckert, S., Liimatta, A., Mahler, H., Bohmann, A., Kozma, G. & Tosin, R. (2025). Digital debating cultures: communicative practices on Reddit. Digital Scholarship in the Humanities, 40(1), 227–240.

Squires, L. (2014). From TV personality to fans and beyond: Indexical bleaching and the diffusion of a media innovation. Journal of Linguistic Anthropology, 24(1), 42–62.

'He's like a god damned laser guided missile': A corpus-assisted analysis of metaphor use in the R/Dementia subreddit

WSP

Gavin Brookes (Lancaster University, UK)

Dementia is a syndrome characterised by a series of diseases that cause impairment in memory, reasoning, perception and communication. Dementia presents one of the greatest medical and social challenges of our time. The way in which linguists, among other

social scientists, have sought to contribute to addressing this challenge is by improving our understanding of how language shapes perceptions of, and attitudes towards, dementia, based for example on analyses of representations found in mass consumption texts like newspaper articles. Much less studied in this regard is the language used by relatives (including relative-carers) of people living with dementia, who in many cases provide the primary means of support for those diagnosed with the syndrome. With this in mind, this study examines the discourse used by such people to represent the person with dementia, their own experiences of the syndrome, and to perform other functions, in the context of the R/Dementia subreddit. This is a large and highly active subreddit that functions primarily as a means for relatives and relative-carers of people with dementia to share their experiences and exchange advice and support.

Based on a purpose-built corpus comprising a sample of a year's worth of posts (2025; including replies) on this subreddit, the analysis focuses in particular on the contributors' use of metaphor. It considers what functions are performed by metaphors in this context, as well as which aspects of dementia and life with dementia are foregrounded by such metaphorical choices (as well as those aspects that might be backgrounded or elided). Comparing the results of this analysis against other – particularly, metaphor-focussed – studies of dementia representation reveals that there are certain kinds of metaphorical constructions that are prevalent across different discourse contexts (e.g., news media, charity campaigns, literary depictions). However, in this context, community members do not simply reproduce such tropes but repurpose them to present a version of life with dementia that is in many ways decidedly different to that which is regularly encountered in such mass audience texts, mostly notably in the sense that it averts the often-fatalistic focus to instead foreground the realities of life with the syndrome in the here-and-now.

Other metaphorical choices observable within the forum can be considered more innovative or at least borrow from contexts that are otherwise sidelined from much mass public discourse, such as from research publications. Such constructions, which become particularly visible when we look beyond what is very frequent, perform important functions within the forum, such as providing humorous accounts of life with a loved one with dementia in a wider process of exchanging support and forging a sense of shared experience. Such metaphors thus play an important role in enabling community members to break away from dominant discourses around dementia in order to construct alternative accounts that convey, perhaps more aptly, the complexities of life with the syndrome (both for the person diagnosed with it, and their loved ones). In light of this, I argue that examining the discourses of these kinds of (virtual) communities, in concert with those of publicly oriented mass consumption texts, can help us to gain a more nuanced view of the lived realities of this global health concern in the here-and-now.

Ban-happy mods and *shadow bans*: A corpus-pragmatic analysis of speech acts on Reddit

WSP

Claudia Lange & Sven Leuckert (Technische Universität Dresden, Germany)

While there is considerable freedom in how users on the social media platform Reddit may express themselves, they are also bound by Reddit's overarching guidelines as well as a set of subreddit-specific rules (Proferes et al. 2021: 1). These rules may be explicit and/or implicit: Many subreddits feature a list of rules on their front page but, in addition, they often also have "specific [unwritten] cultures and norms" (Proferes et al. 2021: 1) that users are expected to adhere to. Sanctioning of rule and norm violations as "a fundamental part of social interaction" (Walz et al. 2024: 137) may take place, for instance, via downvoting posts, which is one of Reddit's main affordances, or by outright banning users. Bans and shadow bans, that is, "the practice of hiding posts from everyone else except for the poster" (Savolainen 2022: 1092), represent a kind of last-resort tool to remove users temporarily or permanently from being able to contribute to a subreddit. Banning represents a controversial practice, since, among other issues, individuals or institutions may transfer "their own cultural biases to the content moderation process" (Thach et al. 2024: 4035).

In order to probe the role of bans and banning practices in the subreddits, we intend to compare the pragmatic functions and usage contexts of mentioning and discussing bans in the two predominantly Indian subreddits r/indianews (ca. 13 million words) and r/IndiaSpeaks (ca. 79 million words) to those in the large gaming subreddit r/leagueoflegends (ca. 1.5 billion words). These subreddits are selected for their size and their different foci, with the two Indian subreddits serving as local discussion forums and the gaming subreddit representing a global, mostly younger demographic with a shared special interest (see Leuckert & Leuckert 2020). Our main research questions are:

(1) Which speech acts do mentions of ban* represent in the subreddits?

(2) In which contexts do users on the subreddits refer to bans?

To address the first question, we develop a bottom-up classification scheme based on the data and apply it to a random sample of 1,000 mentions of ban* per subreddit. To address the second question, we conduct a collocational analysis to identify relevant collocates of ban. Preliminary results reveal that mentioning a ban is frequently part of a threat or an attempt to force accountability (see Walz et al. 2024: 139). However, banning behaviour is often also discussed critically, with users criticising moderation practices as instances of censorship. Overall, the study contributes to a deeper understanding of speech acts on social media (e.g., Dayter 2018) in the context of an important component of many popular platforms.

References

Dayter, D. (2021). Self-praise online and offline: The hallmark speech act of social media? *Internet Pragmatics*, 1(1), 184–203.

Leuckert, S. & Leuckert, M. (2020). Towards a digital sociolinguistics: Communities of practice

on Reddit. In D. Dayter & S. Rüdiger (Eds.), *Corpus Approaches to Social Media* (pp. 15–40). John Benjamins.

Proferes, N., Jones, N., Gilbert, S., Fiesler, C., & Zimmer, M. (2021). Studying Reddit: A systematic overview of disciplines, approaches, methods, and ethics. Social Media + Society, 7(2), 1–14.

Savolainen, L. (2022). The shadow banning controversy: Perceived governance and algorithmic folklore. *Media, Culture & Society*, 44(6), 1091–1109.

Thach, H., Mayworm, S., Delmonaco, D., & Haimson, D. (2024). (In)visible moderation: A digital ethnography of marginalized users and content moderation on Twitch and Reddit. *New Media & Society*, 26(7), 4034–4055.

Walz, L., Joyce, J. B., & Flint, N. (2024). "Facebook's about to know, Karen": Mobilising social media to sanction public conduct. *Internet Pragmatics*, 7(1), 137–160.

Pragmatic strategies and audience response: Advice-giving and upvotes on Reddit

WSP

Rickey Lu (The Hong Kong Polytechnic University, Hong Kong)

This paper explores the pragmatic and interactional dynamics of advice-giving on Reddit, focusing on how advice-givers design their messages for a broad online audience and how these linguistic strategies relate to the platform's upvote system. While advice-giving has been extensively studied in offline contexts, its manifestation in online environments—especially on public, algorithmically mediated platforms like Reddit—remains less understood. I address this gap by examining the types of advice moves used on Reddit, their distribution across various advice topics, and their association with the Reddit audience, as indicated by upvotes.

Employing Swales' move analysis framework and Spencer-Oatey and Kádár's concept of interactional goals, the study analyzes a corpus of 6,621 advice comments from 300 threads in a large advice subreddit. The data were categorized into six main advice topics. Fourteen advice moves were identified, including advice request, agreement, assessment, background, consolation, disclaimer, encouragement, experience sharing, explicit advice, information giving, referral, seeking information, solidarity, and thanks.

Statistical analyses, including Chi-squared tests, Pearson residuals, and Poisson regression modelling, revealed significant associations between advice moves and topics, as well as between moves and upvote counts. The findings highlight how advice-givers adapt their linguistic strategies according to the perceived interactional goals of the advice scenario, which may be transactional (information and problem-solving) or relational (emotional support and rapport-building). Notably, the choice and frequency of advice moves are sensitive to both topic and audience expectations.

This study also applies audience design to the context of Reddit's upvote system, distinguishing between advice-givers, advice-seekers, and the broader audience (auditors and overhearers) who participate indirectly through upvotes. Results indicate misalignment be-

tween the moves utilized by advice-givers and those rewarded by the audience via upvotes. Transactional moves were more frequently associated with increased upvotes, suggesting that the Reddit audience values direct, information-oriented advice over relational, supportive responses. Conversely, relational moves, while prevalent in certain topics, were often linked to fewer upvotes.
This research advances our understanding of advice as a complex, context-sensitive, and role-dependent process in computer-mediated communication. It underscores the importance of considering both language and technological affordances in shaping online advice dynamics. The findings highlight the difficulty for advice-givers to balance topic sensitivity, face concerns, and audience expectations, and point to the need for further research into how move combinations and sequencing may influence audience reception in other digital environments.

Comment type sequences on Reddit: A data-driven functional-pragmatic approach

WSP

Ylva Biri & Aatu Liimatta (University of Helsinki, Finland)

With a focus on corpus-driven methods and text types on Reddit, this paper explores sequences of functional text types in Reddit comment threads. Corpus-driven methods can be used to characterise the functions of Reddit comments and thereby describe their typical communicative functions (e.g. Liimatta 2023). However, Reddit comments are part of a dialogic exchange that provides important analytical context. Rather than focusing on the individual comment in isolation – or on the interaction sequence as a single unit of analysis (e.g. Liimatta 2019; Biber et al. 2021) – we illustrate how interaction can be analysed as a particular sequence of comment types. In this functional-pragmatic approach, individual comments are analysed as performing functions in the social context of interaction between users.
We first classify Reddit comments into comment types based on their functions through statistical analysis of a large-scale dataset of Reddit comments. In practice, we make use of a set of functional lexico-grammatical features adapted from Biber (1988) for online texts; we use latent class analysis (LCA) as a data-driven method to classify the comments into functional types according to the patterns of occurrence of the features in the comments. Through a qualitative analysis of the comments, as well as of the functional profiles of the features characterising each of the comment types, our aim is to find what functional comment types can be identified on Reddit using this approach (RQ1) and, in particular, how do the comment types follow up each other in comment sequences (RQ2): does a comment receive replies with similar functions or are there transitions from one comment type to another? We will explore and analyse these comment type sequences and their pragmatic functions. Furthermore, given that subreddits with different topics have different discussion structures (Yu, Jiang & Dhillon 2024), we also want to find out whether the typical

comment type sequences differ between types of subreddits.
Through a qualitative analysis of the comment type continua and transitions, we explore the types of interactions and activities present on Reddit. Results about how comment types follow each other (RQ2) reveal information about dominant activity types. For instance, in Q&A-based discussions, comments with highly informative functions may be associated with more interactive comments requesting for this information and thanking for the responses (Kiesling et al. 2018). Additionally, we aim to investigate how certain other pragmatic (online) features, such as particular tone indicators (e.g. /jk for joking, /s for sarcasm), are associated with different comment types.
Our pilot study suggests that the method can be used to extract meaningful patterns of comment sequences from Reddit data. For instance, interactive features (e.g. 1st and 2nd person pronouns, private verbs) in particular appear to be meaningful for identifying comment types associated with authorial stance and reader engagement (Hyland 2005).

References

Biber, D., Egbert, J., Keller, D., & Wizner, S. (2021). Towards a taxonomy of conversational discourse types: An empirical corpus-based analysis. Journal of Pragmatics, 171, 20–35. https://doi.org/10.1016/j.pragma.2020.09.018

Biber, D. (1988). Variation across speech and writing. Cambridge University Press. https://doi.org/10.1017/CBO9780511621024

Hyland, K. (2005). Stance and engagement: A model of interaction in academic discourse. Discourse studies, 7(2), 173-192.

Kiesling, S. F., Pavalanathan, U., Fitzpatrick, J., Han, X., & Eisenstein, J. (2018). Interactional stancetaking in online forums. Computational Linguistics, 44(4), 683–718. https://doi.org/10.1162/coli_a_00334

Liimatta, A. (2019). Exploring register variation on Reddit: A multi-dimensional study of language use on a social media website. Register Studies, 1(2), 269–295. https://doi.org/10.1075/rs.18005.lii

Liimatta, A. (2023). Register variation across text lengths: Evidence from social media. International Journal of Corpus Linguistics, 28(2), 202–231. https://doi.org/10.1075/ijcl.20177.lii

Yu, Y., Jiang, J., & Dhillon, P. S. (2024). Characterizing the structure of online conversations across Reddit. Proceedings of the ACM on Human-Computer Interaction, 8(CSCW2), 1–23. https://doi.org/10.1145/3686913

Evaluative practices and persuasion in online debates: a corpus-based study of /r/changemyview

WSP

Daria Dayter (University of Tampere, Finland) & Thomas C. Messerli (University of Basel, Switzerland)

Evaluative practices play a central role in linguistic persuasion as they intersect with key features of persuasive discourse. This overlap can be conceptualised in terms of credibility, stance (DuBois 2007), evaluation (Hunston 2008), appraisal (Martin & White 2005) and positioning (Davies and Harré 2001). For instance, research on liking and persuasion indicates that positive evaluations often help establish common ground and strengthen relational ties, whereas negative evaluations, especially when softened or implicit, can preserve speaker credibility and engagement without overt confrontation (Hyland 2005; Thompson & Hunston 2000). Intersubjective alignment (DuBois 2007) is likewise an important means of fostering solidarity and promoting agreement in public discourse, and it is closely linked to implicitness (Harris et al. 2006).

A key distinction in evaluative practices lies between implicit and explicit evaluation (see, e.g., Bednarek 2009), that is, between using linguistic expressions that encode a stance toward an object and those that imply it and are therefore defeasible. Our study examines pertinent cases of implicit and explicit positive and negative evaluation in the context of persuasion on the subreddit "Change My View" (CMV). Within the digital environment of typically polarised social media discourse, CMV is a community devoted to reasoned debate, where commenters attempt to convince original posters (OPs) to change their views. These persuasive efforts by commenters are regulated by explicit community rules, and they are sanctioned by an emic marker called delta, which is awarded to comments judged, usually by the OP, as successfully persuasive.

In an earlier study of a smaller sample, we examined evaluative practices by commenters in two connected corpora containing delta-awarded responses and non-delta-awarded responses. Expanding on this pilot study and refining and upscaling sampling and coding procedures as well as aspects of the coding scheme (chunk size, operationalising definitions of "explicitness" and "evaluation"), the talk systematically maps evaluative practices across a larger dataset of 500 comments and explores the sequential organisation of evaluative patterns as part of argumentation in comments. The larger sample size and stronger focus on context allow for a more robust assessment of the persuasive impact of evaluation on CMV and in digital discourse more generally.

References

Bednarek, Monika. (2009). Dimensions of evaluation: Cognitive and linguistic perspectives. Pragmatics & Cognition, 17(1), 146–175. https://doi.org/10.1075/pc.17.1.05bed

Davies, Bronwyn, & Harré, Rom. (2001). Positioning The Discursive Production of Selves. In M. Wetherell, S. Taylor, & S. Yates (Eds.), Discourse Theory and Practices: A Reader, (pp. 261-271). Sage.

Du Bois, John W. (2007). The stance triangle. In R. Englebretson (ed.), Stancetaking in discourse: subjectivity, evaluation, interaction (pp. 139–182). John Benjamins.

Harris, Sandra, Karen Grainger and Louise Mullany. (2006). The pragmatics of political apologies. Discourse & Society, 17, 717–736

Hunston, Susan (2008). The evaluation of status in multi-modal texts. Functions of language, 15(1), 64-83.

Hyland, Ken. (2005). Metadiscourse: Exploring Interaction in Writing. Continuum.

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

Thompson, Geoff, & Hunston, Susan eds. (2000). Evaluation in Text: Authorial Stance and the Construction of Discourse. Oxford University Press.

Workshop : Corpus and Computational Linguistics meet Fake News, Mis- and Disinformation and Large Language Models Room E314 • 9:00–12:30

Conveners: Silje Susanne Alvestad (University of Oslo) • Nele Põldvere (University of Oslo)

This workshop will take a corpus- and computational-linguistics perspective on fake news and related phenomena, where fake news is defined along the axes of veracity and honesty, giving rise to three types: 1) false but honest news, such as errors, which corresponds to misinformation; 2) false and dishonest news, such as lies; and 3) true but dishonest news, in which crucial pieces of information may be omitted (so as to fit a certain narrative, as seen, arguably, in propaganda), or in which true information may be taken out of context. The workshop will shed light on the rising societal challenge posed by information disorders from a corpus- and computational-linguistics perspective.
Contacts: s.s.alvestad@ilos.uio.no, nele.poldvere@ilos.uio.no

Presentations:

Corpus and Computational Linguistics Meet Fake News, Mis- and Disinformation and Large Language Models

WSP

Silje Susanne Alvestad & Nele Põldvere (University of Oslo, Norway)

This workshop will take a corpus- and computational-linguistics perspective on fake news and related phenomena, where fake news is defined along the axes of veracity and honesty, giving rise to three types: 1) false but honest news, such as errors, which corresponds to misinformation; 2) false and dishonest news, such as lies; and 3) true but dishonest news, in which crucial pieces of information may be omitted (so as to fit a certain narrative, as seen, arguably, in propaganda), or in which true information may be taken out of context. Fake news types 2) and 3) involve an intention to deceive and so overlap with typical definitions of disinformation (see Grieve & Woodfield, 2023)
Fake news and related information disorders can be harmful to our societies. Specifically, when we change our beliefs and subsequent behaviour based on false or misleading information it can harm our health and lives, sow distrust (Funk et al., 2023), and disrupt election processes (Jamieson 2018). Now, the societal challenge posed by information disorders is amplified by the rapid development within generative AI, exemplified by Large Language Models (LLMs), with the launch of OpenAI's ChatGPT in November 2022 as a significant milestone. The output of LLMs depends on their training data, which can contain inaccuracies and biases. As a result, these models may unintentionally spread mis- or disinformation (Brandtzæg et al., 2023). They can also produce "hallucinations"—convincing but false statements (Spitale et al., 2023)—or partly incorrect content due to unreliable sources (Chen et al., 2023). This blend of fabricated and biased information makes it difficult

to ensure the accuracy of online content (Buchanan et al., 2021). Moreover, LLMs hold the potential to generate misleading or false information at scale and at a quality that makes it indistinguishable from similar content authored by humans. Controlled experiments show that LLM-generated messages can change policy attitudes, at times matching or surpassing human levels of persuasiveness (Bai et al., 2025; Salvi et al., 2025). Research has shown that people find it more difficult to identify disinformation produced by AI than similar content produced by humans (Zhou et al., 2023), and in simulated news recommendation systems, researchers have found a new phenomenon referred to as "truth decay", by which genuine news increasingly falls behind LLM-generated mis- and disinformation in visibility and ranking. This shift happens because LLM-generated content typically shows lower perplexity, making it appear more fluent and familiar. As a result, such content often receives higher recommendation scores and greater visibility (Hu et al., 2025). This dynamic has serious implications for the spread of mis- and disinformation, since increased exposure can boost perceived credibility through the illusory truth effect. All of this highlights the need for effective identification and verification systems. We believe that especially corpus and computational linguists should recognize the urgency of the moment and hereby be invited to act.

Against this background, our workshop will shed light on the rising societal challenge posed by information disorders from a corpus- and computational-linguistics perspective. We welcome abstracts from both branches of linguistics that examine LLM-generated as well as human-authored fake news and other types of misleading or false information in English, in comparison or separately, and similarly for various types of LLMs. We ask questions including, but not limited to, what the linguistic features are of such information disorders, whether the disorders can be identified based on such features, whether the features have changed, and are changing, over time, what the capabilities and limitations of various LLMs are in the context of producing and disseminating misleading information, whether the LLMs have any fingerprint in the context of mis- and disinformation, and how to develop a best practice for linguistic investigations of LLM output. Abstracts can address theoretical as well as methodological questions, take a comparative or case-focused approach, and examine human-authored or LLM-generated text, or both.

Abstracts of a maximum of 400 words, excluding references, should be sent to the workshop organisers by December 15, with notification of outcome on December 20. Authors of accepted abstracts will be invited to present their research at the pre-conference workshop of ICAME47 on Tuesday, 26 May 2026.

References

Bai, H., Voelkel, J., Muldowney, S., Eichstaedt, J., & Willer, R. (2025). LLM-generated messages can persuade humans on policy issues. Nature Communications, 16, Article 61345. https://doi.org/10.1038/s41467-025-61345-5

Brandtzaeg, P. B. (2023). "Good" and "Bad" Machine Agency in the Context of Human-AI Communication: The Case of ChatGPT. In International Conference on Human-Computer Interaction, (pp. 3–23). Springer Nature Switzerland.

Buchanan, B., Lohn, A., Musser, M. & Sedova, K. (2021). Truth, lies and automation. How language models can change disinformation. Center for Security and Emerging Technology, May 2021. https://doi.org/10.51593/2021CA003

Chen, C. & Shu, K. (2023). Combating misinformation in the age of LLMs: Opportunities and challenges. https://doi.org/10.48550/arXiv.2311.05656

Funk, A., Shahbaz, A., & Vesteinsson, K. (2023). Freedom on the Net 2023. The repressive power of artificial intelligence. Freedom House report. https://freedomhouse.org/sites/default/files/2024-10/FOTN2023Final24.pdf

Grieve, J., & Woodfield, H. (2023). The Language of Fake News. Cambridge Elements in Forensic Linguistics. Cambridge University Press

Hu, B., Sheng, Q., Cao, J., Li, Y., & Wang, D. (2025). LLM-generated fake news induces truth decay in news ecosystem: A case study on neural news recommendation. In Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 435–445). Association for Computing Machinery. https://doi.org/10.1145/3726302.3730027

Jamieson, K. H. (2018). Cyberwar: How Russian hackers and trolls helped elect a president. What we don't, can't, and do know. Oxford University Press.

Salvi, F., Horta Ribeiro, M., Gallotti, R., & West, R. (2025). On the conversational persuasiveness of GPT-4. Nature human behaviour, 9(8), (pp. 1645–1653). https://doi.org/10.1038/s41562-025-02194-6

Spitale, G., Biller-Andorno, N., & Germani, F. (2023). AI model GPT-3 (dis)informs us better than humans. doi: 10.1126/sciadv.adh1850.

Zhou, J., Zhang, Y., Luo, Q., Parker, A. G., & Choudhury, M. D. (2023). Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions. In CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3544548.3581318

Stance expressions in AI-generated vs. human-written fake news

WSP

Sophie Llewellyn (University of Oslo, Norway)

ChatGPT (Chat Generative Pre-trained Transformer) is a chatbot developed by OpenAI that uses artificial intelligence (AI) to produce human-like text. Whilst the capabilities of the chatbot are vast, one concern is its potential exploitation to generate fake news. Previously, the fabrication and distribution of fake news was a labour-intensive process requiring a significant amount of time and money. However, AI chatbots now simplify this process, as they can produce proficient and credible-sounding articles, often indistinguishable from human-written content, within seconds (Kreps et al. 2020). Consequently, the accessibility and abilities of AI chatbots could lead to an unprecedented level of fake news content.
To help combat this risk, more research is needed to be able to identify AI-generated fake news, including linguistic analyses. This corpus-based study is inspired by Grieve and Woodfield's (2023) research on the language of human-written fake news and builds on their

findings related to conviction by investigating stance expressions in AI-generated fake news. More specifically, this study seeks to answer the following research question:
How does the use of epistemic stance adverbials and stance adjectives controlling complement clauses compare in fake news generated by ChatGPT and fake news written by humans?
To address this research question, two corpora of fake news articles were used – the purpose-built ChatGPT Fake News Corpus and a subsection of the Fakespeak News Corpus, each containing 200 fake news articles on identical topics. (Currently, the Fakespeak News Corpus is not publicly available, but for information about the corpus, see Põldvere et al. 2023.) These corpora facilitated the analysis of 40 different adverbial types (e.g. certainly, perhaps, truly), and 18 different adjective types controlling that and to-complement clauses (e.g. it is clear that, it is likely to), which were identified based on research by Biber et al. (1999) and Biber (2006).
Findings indicate that stance adverbials and stance adjectives controlling to-complement clauses occur significantly more frequently in human-written fake news than AI-generated news. In contrast, stance adjectives controlling that-complement clauses occur more frequently in AI-generated fake news than human-written fake news. Results also find that unlike the varied use in the human-written articles, stance expressions regularly occur in fixed patterns the AI-generated articles, particularly within quoted speech or article endings. Altogether, findings from this study indicate that human-written fake news articles use stance expressions more frequently and with more variety than AI-generated fake news articles.

References

Biber, Douglas. 2006. University Language: A Corpus-based Study of Spoken and Written Registers. Amsterdam: John Benjamins Publishing.

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

Grieve, Jack, and Helena Woodfield. 2023. The Language of Fake News. Cambridge: Cambridge University Press.

Kreps, Sarah R., Miles McCain, and Miles Brundage. 2020. "All the News That's Fit to Fabricate: Ai-Generated Text as a Tool of Media Misinformation." SSRN Electronic Journal, (September). https://doi.org/10.2139/ssrn.3525002.

Põldvere, Nele, Zia Uddin, and Aleena Thomas. 2023. "The PolitiFact-Oslo Corpus: A New Dataset for Fake News Analysis and Detection." Information 14, no. 12 (November). https://doi.org/10.3390/info14120627

Modes of persuasion: Can LLMs be deceived into trusting disinformation?

WSP

Marina Ernst & Frank Hopfgartner (University of Koblenz, Germany)

While misinformation is not a novel phenomenon, advances in information technology have amplified its generation and spreading. Prior work indicates an increase of misinformation over recent decades (Broda, Strömbäck, 2024), fostering what is commonly described as a "post-truth" era (Lewandowsky et al., 2017).

Generative AI technologies have significantly transformed the misinformation and disinformation ecosystem, affecting the generation, spread, and detection of false content (Bontridder, Poullet, 2021). At the same time, Gen AI in general, and Large Language Models in particular, can be leveraged to combat misinformation and disinformation, with prior work exploring their effectiveness as fact-checking and debunking agents (Jiang et al., 2024; Pelrine et al., 2023). However, the reasoning behind the LLMs decisions often remains a black box, blemishing transparency and leaving open the opportunity to deceive them into misclassifying potentially harmful content.

As disinformation is distinguished from misinformation by its intentional, deceptive, and goal-directed nature, it often relies on persuasive tactics (Hameleers, 2022). Therefore, it often employs traditional persuasion modes, including rhetorical appeals (ethos, pathos, and logos) (Amos et al., 2022). In this study we suggest applying those three dimensions as a framework to shed light on the effect of rhetorical and linguistic characteristics of English text on the process of LLM-based disinformation detection.

The dataset to be used consists of 367 news articles from internet outlets in English, labeled as "disinformation" or "reliable" by the team of debunking experts according to the methodology described by Modzelewski et al. (2024). As an initial step, three state-of-the-art LLMs will be tasked to detect disinformation in zero-shot settings. Next, each text will be analyzed using the three persuasion modes. Finally, quantitative analysis will be applied to evaluate the effect of each characteristic on LLM decisions and, consequently, whether manipulating these characteristics could lead LLMs to "trust" disinformative content.

To assess the potential impact of each mode of persuasion, they are presented in a quantifiable manner. For Pathos, sentiment analysis is use to quantify emotional content in text, which serves as a proxy for emotional appeal (Evgrafova et al., 2024). Ethos as rhetorical credibility construction is measured via the presence of authority-signaling linguistic cues (e.g., references to experts, studies, or institutions) (Herman, 2022). We operationalize logos as evaluable propositional content and use a transformer-based natural language inference model to identify claims expressing factual assertions amenable to logical or evidential evaluation (Ruiz-Dolz et al., 2021).

References

Amos, C., Zhang, L., King, S., & Allred, A. (2022). Aristotle's modes of persuasion and valence effects on online review trustworthiness and usefulness. Journal of Marketing Communications,

28(4), 360–391. doi:10.1080/13527266.2021.1881806

Bontridder, N., & Poullet, Y. (2021). The role of artificial intelligence in disinformation. Data & Policy, 3, e32. doi:10.1017/dap.2021.20

Broda, E., & Strömbäck, J. (2024). Misinformation, disinformation, and fake news: lessons from an interdisciplinary, systematic literature review. Annals of the International Communication Association, 48(2), 139–166. doi:10.1080/23808985.2024.2323736

Evgrafova, N., Hoste, V., & Lefever, E. (2024, May). Analysing Pathos in User-Generated Argumentative Text. In H. Afli, H. Bouamor, C. B. Casagran, & S. Ghannay (Eds), Proceedings of the Second Workshop on Natural Language Processing for Political Sciences @ LREC-COLING 2024 (pp. 39–44). Retrieved from https://aclanthology.org/2024.politicalnlp-1.5/

Hameleers, M. (10 2022). Disinformation as a context-bound phenomenon: toward a conceptual clarification integrating actors, intentions and techniques of creation and dissemination. Communication Theory, 33(1), 1–10. doi:10.1093/ct/qtac021

Herman, T. (2022). Ethos and Pragmatics. Languages, 7(3). doi:10.3390/languages7030165

Jiang, B., Tan, Z., Nirmal, A., & Liu, H. (2024). Disinformation Detection: An Evolving Challenge in the Age of LLMs. In Proceedings of the 2024 SIAM International Conference on Data Mining (SDM) (pp. 427–435). doi:10.1137/1.9781611978032.50

Lewandowsky, S., Ecker, U. K. H., & Cook, J. (2017). Beyond Misinformation: Understanding and Coping with the "Post-Truth" Era. Journal of Applied Research in Memory and Cognition, 6(4), 353–369. doi:10.1016/j.jarmac.2017.07.008

Modzelewski, A., Da San Martino, G., Savov, P., Wilczyńska, M. A., & Wierzbicki, A. (2024, November). MIPD: Exploring Manipulation and Intention In a Novel Corpus of Polish Disinformation. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 19769–19785). doi:10.18653/v1/2024.emnlp-main.1103

Pelrine, K., Imouza, A., Thibault, C., Reksoprodjo, M., Gupta, C., Christoph, J., ... Rabbany, R. (2023, December). Towards Reliable Misinformation Mitigation: Generalization, Uncertainty, and GPT-4. In H. Bouamor, J. Pino, & K. Bali (Eds), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 6399–6429). doi:10.18653/v1/2023.emnlp-main.395

Ruiz-Dolz, R., Alemany, J., Barberá, S. M. H., & García-Fornes, A. (2021). Transformer-Based Models for Automatic Identification of Argument Relations: A Cross-Domain Evaluation. IEEE Intelligent Systems, 36(6), 62–70. doi:10.1109/MIS.2021.3073993

Fake and real, beyond binary classification in fake news detection: Challenges and solutions with topic modelling techniques

WSP

Stefano Sbalchiero, Alessandro Meneghini & Arjuna Tuzzi (University of Padova, Italy)

The detection of fake news is typically approached as a binary classification problem, trying to identify "real" from "fake" content. This framing, however, fails to capture the nuances of misinformation, where factual and fake information are combined to enhance their plausibility and persuasive power. To move beyond this binary paradigm, this study introduces

the concept of a gray area in misinformation, conceptualized as semantic spaces where real and fake elements are discoursively overlapping. The classical approach to fake news detection treats the problem as a binary classification task in which the central question is "What is true?" Machine learning systems are evaluated - and improved - based on their ability to categorise items as either "real" or "fake." While effective for many applications, this binary framing fails to capture the complex and nuanced nature of misinformation. The boundaries of what constitutes fake news are often ambiguous and overlap with related phenomena such as misinformation, disinformation, propaganda, satire, hoaxes, and conspiracy theories. Consequently, the term "fake news" has become a catch-all label encompassing content created for profit, content designed to discredit others, factual information distorted through selective framing, and even news that simply conflicts with someone's beliefs. More recent research and the work of professional fact-checkers show that many deceptive items contain mixtures of true and false elements - such as accurate data embedded in misleading context - making them difficult to classify definitively. This phenomenon has been described as the "grey area" or "grey fake news." The present study adopts a conceptual shift. Instead of asking "What is true?" we ask "What is involved in the construction of a truth?" Our goal is not to categorise entire documents as real or fake, but to measure the probability that individual pieces of news contain both truthful and misleading elements. To do so, we adopt the concept of "noise" not as an imperfection of algorithms but as a structural component of information itself. We employ Structural Topic Modeling (STM) over an English corpus composed of COVID-19 news headlines to map this landscape. Our analysis reveals a spectrum of topics, ranging from those strongly polarized towards factual reporting to those dominated by conspiratorial and politicized frames. Most significantly, we identify a set of noisy topics, thematic clusters centered on issues like public health guidance and disease characteristics that are not associated to either real or fake labels. The specific words defining these noisy topics explain why certain types of misleading content are particularly challenging to detect and debunk, as they are semantically embedded within the same informational frameworks as legitimate news. This finding argues for a fundamental shift in detection strategies, from classifying entire documents to identifying and monitoring these ambiguous topics where the line between fact and falsehood is most frequently blurred.

References

Ahmed, A. A. A., Aljarbouh, A., Donepudi, P. K., Choi, M. S.: Detecting Fake News Using Machine Learning: A Systematic Literature Review. arXiv:2102.04458 (2021).

Allcott, H., Gentzkow, M.: Social Media and Fake News in the 2016 Election. Journal of Economic Perspectives. 31(2), 211-236 (2017).

Asubiaro, T. V., Rubin, V. L.: Comparing features of fabricated and legitimate political news in digital environments (2016-2017). Proceedings of the Association for Information Science and Technology. 55(1), 747-750 (2018).

Bhavani, A., Santhosh Kumar, B.: A Review of State Art of Text Classification Algorithms. 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), 1484-1490

(2021). https://doi.org/10.1109/ICCMC51019.2021.9418262

Blei, D., Carin, L., Dunson, D.: Probabilistic Topic Models. IEEE Signal Processing Magazine. 5563111 (2010). https://doi.org/10.1109/MSP.2010.938079

Blei, D. M.: Latent Dirichlet Allocation. Journal of Machine Learning Research. 3 (2003).

Deepak, P., Tanmoy, C., Cheng, L., Santhosh Kumar, G. (eds.): Data Science for Fake News. Surveys and Perspectives. Springer Nature Switzerland (2021).

Kaliyar, R. K., Goswami, A., Narang, P.: DeepFakE: Improving fake news detection using tensor decomposition-based deep neural network. Journal of Supercomputing. 77, 1015-1037 (2020).

Patwa, P., et al.: Fighting an Infodemic: COVID-19 Fake News Dataset. In: Chakraborty, T., Shu, K., Bernard, H. R., Liu, H., Akhtar, M. S. (eds.) Combating Online Hostile Posts in Regional Languages during Emergency Situation. Springer International Publishing. 1402, 21-29 (2021). https://doi.org/10.1007/978-3-030-73696-5_3

Reimers, N., Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv:1908. 10084 (2019). http://arxiv.org/abs/1908.10084

Roberts, M. E., et al.: Structural topic models for open-ended survey responses. American Journal of Political Science. 58(4), 1064-1082 (2014).

Rodríguez-Ferrándiz, R.: An overview of the fake news phenomenon: From untruth-driven to post-truth-driven approaches. Media and Communication. 11(2), 15-29 (2023).

Rudloff, J. P., Hutmacher, F., Appel, M.: Post-truth epistemic beliefs rooted in the dark factor of personality are associated with higher COVID-19 vaccination refusal. Scientific Reports. 13(1), 4254 (2023).

Testing Community-level interventions in online echo chambers using LLM agents

WSP

Fabio Carrella (University of Campinas, Brazil)

In this work, we use LLM-based agent simulations conducted in English, in which both community members and intervention bots are synthetic personas, to model anti-vaccine communities that reproduce key structural and dynamical features commonly observed in real online settings. In control simulations without interventions, agents display strong attitudinal homophily in network formation, preferentially forming connections with others who share similar anti-vaccine stances, and the resulting networks exhibit increasing inequality in user popularity as they grow. Information diffusion is similarly organized: agents predominantly repost attitude-aligned content, reinforcing group boundaries, while attention is unequally concentrated, with a small fraction of posts accounting for the majority of reposts and most content receiving little or no engagement.
We then introduce a small fraction of pro-vaccine bots of two types: generic bots and empathetic, attitude-targeted (ERI-based) bots, to test whether attitude-targeted messaging can alter these digital collective patterns. While bot presence does not substantially reduce homophily or overall inequality, both generic and ERI-based bots attenuate network growth,

reducing the likelihood that the authors of posts they engage with gain followers. In contrast, reductions in reposting are specific to ERI-based bots, with generic bots showing no significant effect on content diffusion. Importantly, this attenuation persists for ERI-based bots even when their comments appear alongside agent comments, weakening the usual amplification produced by social interaction. However, these intervention effects coexist with negative community reception: bot-generated comments are generally met with resistance, though empathetic, attitude-targeted bots receive more favourable reactions than generic pro-vaccine bots. Together, these results highlight both the potential and the limits of scaling individual-level engagement strategies to simulated community-level dynamics, revealing trade-offs between misinformation reduction, growth suppression, and community reception.

References

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2016). The spreading of misinformation online. Proceedings of the national academy of Sciences, 113(3), 554-559.

Brugnoli, E., Cinelli, M., Quattrociocchi, W., & Scala, A. (2019). Recursive patterns in online echo chambers. Scientific reports, 9(1), 20118.

Horta Ribeiro, M., Jhaver, S., Zannettou, S., Blackburn, J., Stringhini, G., De Cristofaro, E., & West, R. (2021). Do platform migrations compromise content moderation? evidence from r/the_donald and r/incels. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2), 1-24.

Russo, G., Verginer, L., Ribeiro, M. H., & Casiraghi, G. (2023, June). Spillover of antisocial behavior from fringe platforms: The unintended consequences of community banning. In Proceedings of the international AAAI conference on web and social media (Vol. 17, pp. 742-753).

Holford, D., Schmid, P., Fasce, A., & Lewandowsky, S. (2024). The empathetic refutational interview to tackle vaccine misconceptions: Four randomized experiments. Health Psychology, 43(6), 426.

Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. Science, 385(6714), eadq1814.

Workshop : ICE Corpora in the Age of AI Room E114 (afternoon session)
• 14:00–17:30

Conveners: Ulrike Gut (University of Münster) • Stella Neumann (RWTH Aachen University) • Gerold Schneider (University of Zurich)

The International Corpus of English (ICE) has been a cornerstone of World Englishes research for three decades. This workshop examines the challenges and opportunities facing ICE-family corpora in the age of AI: updated text category designs, automated transcription, phonological annotation, and the question of what "quality" means for a reference corpus when AI tools are both a resource and a research object. Contact: gut@uni-muenster.de

Presentations:

From annotation to automation: Leveraging AI to build phonological corpora of West African Englishes

WSP

Philipp Meer & Polina Kashkarova (University of Münster, Germany)

For the systematic study of phonological properties of varieties of English, large speech corpora are necessary. Recent technological advances such as automatic speech recognition (ASR) tools can facilitate and speed up their compilation. In this talk, we report on ongoing work to enrich ICE Nigeria and ICE Ghana with phonetic and phonological annotations and to compile a phonological corpus of Cameroonian English for the study of vowel and consonantal differences between West African Englishes. For the existing ICE Nigeria corpus (Wunder et al. 2010), which already contains time-aligned orthographic transcriptions, phonemic annotations are added; for ICE Ghana, time-aligned orthographic transcriptions are created before adding phonemic annotations; the Cameroonian corpus is newly compiled (modelled after the spoken part of the ICE corpora, Greenbaum & Nelson 1996) and phonologically annotated.

First, we present how we create orthographic transcriptions using ASR and speaker diarization with WhisperX (Bain et al. 2023) for the Cameroonian data. The output is manually checked for transcription accuracy, correct speaker identification and utterance-level alignment in ELAN. Second, we show how we augment all three corpora with phonemic annotations by creating time-aligned phonemic transcriptions with FAVE-align (Rosenfelder et al. 2014) and the Montreal Forced Aligner (McAuliffe et al. 2017). Both aligners have been shown to perform well in the segmentation of different varieties of English compared with manual human alignment, even slightly more so than MAUS (Gonzalez et al. 2020, MacKenzie & Turton 2020, Meer 2020). Automatic alignment is followed by manual correction by phonetically trained human transcribers in Praat.

The opportunities and challenges of using these automatic speech processing tools for the

compilation of phonological corpora will be discussed. In addition, we will demonstrate some examples of how the annotated corpora can be used to analyze the phonetics and phonology of West African Englishes.

References

Bain, Max, Jaesung Huh, Tengda Han & Andrew Zisserman. 2023. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. Proceedings of Interspeech. ISCA: Dublin.

Gonzalez, S., Grama, J. & Travis, C. E. 2020. Comparing the performance of forced aligners used in sociophonetic research. Linguistics Vanguard, 6(1), 1-13.

Greenbaum, S., & Nelson, G. (1996). The International Corpus of English (ICE) project. World Englishes, 15, 3–15. https://doi.org/10.1111/j.1467-971x.1996.tb00088.x

McAuliffe, M., Fatchurrahman, M. R., Feiteng, GalaxieT, NTT123, Gulati, A., Coles, A., Kong, C., Veaux, C., Eren, E., Gritskevich, E., Thor, G., Mishra, H., Ogasawara, H., Fruehwald, J., Maria, Potrykus, P., Jung, S., Sereda, T.,...彭震东. (2025). Montreal Forced Aligner (Version 3.3.6) [Software]. Zenodo. https://doi.org/10.5281/zenodo.17203348

MacKenzie, L. & Turton, D. 2020. Assessing the accuracy of existing forced alignment software on varieties of British English. Linguistics Vanguard 6(1), 1-14.

Meer, P. 2020. Automatic alignment for New Englishes: Applying state-of-the-art aligners to Trinidadian English. The Journal of the Acoustical Society of America 147(4), 2283- 2294.

Rosenfelder, I., Fruehwald, J., Evanini, K., Seyfarth, S., Gorman, K., Prichard, H. & Yuan, J. 2014. "FAVE (Forced Alignment and Vowel Extraction). Program Suite v1.2.2" https://github.com/JoFrhwld/FAVE

Wunder, E.-M., Voormann, H. & Gut, U. 2010. The ICE Nigeria corpus project: Creating an open, rich and accurate corpus. ICAME Journal 34, 78-88.

ICE Transformers

WSP

Henning Schreiber & Ismail Afolabi (Hamburg University, Germany)

The International Corpus of English (ICE; Greenbaum, 1996; Greenbaum & Nelson, 1996) has been instrumental in revealing important insights about World Englishes through its extensive, balanced, and comparable collections from 15 countries, which are continuously expanding. The tremendous changes brought about by the digital age and globalization have also impacted English corpus linguistics. Although ICE has been successful, it requires a critical reassessment due to factors such as alternative methods (e.g., the "web as corpus" approach), data collection via web scraping, and new standards for machine readability and multimodality in linguistic data–along with a need for richer sociolinguistic metadata
In this presentation, we will report on the use of AI tools for the rapid and efficient processing of field recordings from Lagos, Nigeria. Calbert & Roll (2024) evaluate OpenAI's Whisper ASR, demonstrating that its transcription accuracy varies significantly by accent and speaker traits, with consistently strong performance for major native-English accents but noticeably

higher error rates for many non-native and underrepresented varieties. We will discuss potential strategies for compiling ICE resources with a particular focus on including non-elite varieties, as well as identifying heavy L1 interference and creole continua. Additionally, we will address the extraction of sociolinguistic metadata from verbally reported linguistic biographies using Large Language Models.

References

Calbert G. & Roll, N. (2024). Evaluating OpenAI's Whisper ASR: Performance analysis across diverse accents and speaker traits. JASA Express Lett 4(2): https://doi.org/10.1121/10.0024876

Greenbaum, S. (1996). Comparing English Worldwide. Oxford University Press. Greenbaum, S., & Nelson, G. (1996). The International Corpus of English (ICE) project. World Englishes, 15(1), 3–15. https://doi.org/10.1111/j.1467- 971x.1996.tb00088.x

Schreiber, H. & I. Afolabi. (2025). Forschungsgruppe FOR 5728 - CODILAC–Lagos. https://gepris.dfg.de/gepris/projekt/549355286

ICE– Corpora and AI

WSP

John Kirk (University of Vienna, Austria)

This paper will show how the AI tools Gemini and Claude can not only annotate an ICE-corpus pragmatically but also provide, on the basis of that annotation, a thorough and insightful analysis. The case in question is ICE-Ireland, and the use of the SPICE-Ireland Annotation Scheme (Kallen & Kirk 2012, Kirk 2016). The annotation of texts from SPICE-Ireland will be replicated and the outcomes compared with the original SPICE annotation. The results are highly impressive as they go well beyond the achievements of the original scheme.

References

Kallen, J. L. & J. M. Kirk. 2012. *SPICE-Ireland: A User's Guide*. Belfast: Cló Ollscoil na Banríona.

Kirk, J. M. 2016. 'The Pragmatic Annotation Scheme of the SPICE-Ireland Corpus'. In *International Journal of Corpus Linguistics*. 21(3), pp. 299–323.

AI-enhanced and smartphone-based compilation of spontaneous speech: The creation of the Corpus of Spoken Maldivian English (CoSpoMalE)

WSP

Aishath Suad (Maldives National University, Maldives), Tobias Bernaisch (Justus Liebig University Giessen, Germany), Julia Degenhardt (Augsburg University, Germany), Barbara Güldenring (Justus Liebig University Giessen, Germany) & Eliane Lorenz (Justus Liebig University Giessen, Germany)

Contemporary spoken corpus data of varieties of English are rare. With a focus on World

Englishes, this lack of spoken corpus resources is particularly lamentable since the individuality of varieties of English manifests itself in "the subconscious set of conventions regulating the norm level of speech habit, of what is normally said"(Schneider 2007: 92).
The reason for this lack of contemporary spoken corpora of World Englishes is that their compilation has traditionally been particularly labour-intensive, especially when corpus designs like that of the International Copus of English (cf. Greenbaum 1991; Greenbaum & Nelson 1996) require the integration of delicate text categories like legal cross-examinations. Yet, the collection of speech samples on site also offers the opportunity of gathering detailed sociolinguistic information – either through questionnaires before or conversation prompts during recordings–that is generally unavailable in written megacorpora (cf. e. g. Leimgruber et al. 2022 for primarily sociolinguistic data collection through speech recordings in the United Arab Emirates).
Generative artificial intelligence (GenAI) and more specifically speech-to-text models have the potential to lastingly revitalise the compilation of spoken corpora. In this light, the paper at hand addresses the following research questions:
In what ways does GenAI facilitate the compilation process of spoken corpora? How can the collection of spoken corpus data be combined with sociolinguistic queries?
How can the corpus compilation process be made convenient for informants and researchers?
Said research questions are discussed with regard to the compilation of the Corpus of Spoken Maldivian English (CoSpoMalE), the first spoken corpus of Maldivian English. The corpus compilation process entailed 1) obtaining consent and detailed sociobiographic information from the informants, 2) the recording of the informants'conversations and 3) the AI-supported transcription of these recordings. Corpus compilation steps 1) and 2) were completed on smartphones–informants filled in consent and sociobiographic forms on LimeSurvey (Limesurvey GmbH n. d.) on their phones, they downloaded instructions for the conversations and the recordings were made through Android and iOS apps that can save audio in high-quality .wav format. The recordings in CoSpoMalE represent face-to-face discussions between two Maldivian speakers about 16 pre-selected topics that, initially, reduce unfavourable effects of the observer's paradox (Labov 1972: 209) through questions about emojis and social media, but later address sociolinguistic topics such as code-switching and attitudes towards varieties. This sociolinguistic-corpus-discussion approach allows for structural as well as for content analyses of the resulting corpus data. For step 3, initial, AI-assisted transcripts have been created through the open-source software noScribe (Dröge 2024) and refined through manual edits by Maldivian speakers of English.
This paper outlines a cost-effective workflow for the creation of spoken corpora and shows how AI-driven tools can transform corpus compilation from a labour-intensive practice into a community-based process. This type of corpus compilation facilitates the empirical sociolinguistic and structural representation of World Englishes like Maldivian English that have not yet gotten the attention they deserve.

References

Dröge, K. (2024). noScribe. AI-powered audio transcription (Version 0.6.2) [Computer software].

Greenbaum, S. (1991). The development of the International Corpus of English. In K. Aijmer, & B. Altenberg (Eds.), English Corpus Linguistics: Studies in Honour of Jan Svartvik (pp. 83–92). Longman.

Greenbaum, S., & Nelson, G. (1996). The International Corpus of English (ICE) project. World Englishes, 15, 3–15.

Labov, W. (1972). Sociolinguistic patterns. University of Pennsylvania Press. Limesurvey GmbH. (n.d.). LimeSurvey: an open source survey tool. LimeSurvey GmbH.

Leimgruber, J., Al-Issa, A., Lorenz, E., & Siemund, P. (2022). Managing and investing in hybrid identities in the globalized United Arab Emirates. Journal of Language, Identity and Education, 23, 955–972.

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

ICE21: Suggestions for an updated high-quality corpus capturing the diversity of English

WSP

Stella Neumann (RWTH Aachen University, Germany)

The International Corpus of English as an overarching initiative in corpus linguistics (Greenbaum, 1996) is unparallelled in the range of registers covered across a broad range of varieties of English. While there may be much larger corpora, none of them comes close to covering the range of variation captured systematically by the ICE Corpus across its individual components. At the same time, the ICE Corpus in its present state has some serious weaknesses that constrain its usefulness for the above-mentioned research questions. Among the most problematic aspects are challenges in collecting texts that match text categories across largely different cultures, the focus on ”educated” English, which constrains the coverage of vernacular language use, and technical problems such as mark-up issues, diverging formats across components, etc.

The International Corpus of English as an overarching initiative in corpus linguistics (Greenbaum, 1996) is unparallelled in the range of registers covered across a broad range of varieties of English. While there may be much larger corpora, none of them comes close to covering the range of variation captured systematically by the ICE Corpus across its individual components. At the same time, the ICE Corpus in its present state has some serious weaknesses that constrain its usefulness for the above-mentioned research questions. Among the most problematic aspects are challenges in collecting texts that match text categories across largely different cultures, the focus on ”educated” English, which constrains the coverage of vernacular language use, and technical problems such as mark-up issues, diverging formats across components,etc.

One way for corpus linguistics to remain relevant - and to answer meaningful research questions - in the age of ubiquitous text generation, is to tackle the diversity of language in

its entirety and to do so in a transparent way.
In this perspective, accounting for how individuals with their specific social and regional backgrounds use language (that may or may not involve text generation) against the backdrop of population-level (register/variety) patterns is a central question - at this level of specificity. This means that the same amount of comprehensive metadata on every language user contributing (to) a text in a corpus is of the essence, as demonstrated by recent best practice examples such as the Spoken BNC2014 (Love et al., 2017). Moreover, citizen-science approaches to data collection as demonstrated by the Spoken BNC2014 and LANA-CASE (Hanks et al., 2024) can ensure better coverage of vernacular language use within the respective speech communities. The targeted register coverage needs to be flexibilised to reflect differences in the cultural contexts of the respective speech communities. This paper makes a case for a new corpus collection campaign in the spirit of the International Corpus of English that maintains the careful curation of texts across highly diverse contexts, while overcoming the above mentioned issues.

References

Biber, D. (1995). Dimensions of register variation. Cambridge University Press.

Biber, D. (2012). Register as a predictor of linguistic variation. Corpus Linguistics and Linguistic Theory, 8(1), 9–37.

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

Halliday, M. A. K., McIntosh, A., & Strevens, P. (1964). The Linguistic Sciences and Language Teaching. Longman.

Hanks, E., Clarke, I., Brookes, G., Brezina, V., Baker, P., Reppen, R., Biber, D., Larsson, T., Egbert, J., McEnery, T., & 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), Article 2. https://doi.org/10.32714/ricl.12.02.03

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. International Journal of Corpus Linguistics, 22(3), 319–344. https://doi.org/10.1075/ijcl.22.3.02lov

Neumann, S. (2014). Contrastive register variation: A quantitative approach to the comparison of English and German. de Gruyter Mouton.

Balancing backward comparability and register adequacy: Towards an ICE 2.0 text category design in the age of AI

WSP

Robert Fuchs (University of Bonn, Germany)

The International Corpus of English (ICE) has provided a uniquely comparable, register-balanced set of corpora for varieties of English worldwide (Greenbaum 1996; Kirk & Nelson 2018). Each component contains around one million words sampled across 32 spoken and

written text categories. 35 years since the inception of the ICE project, this 'small and tidy' design contrasts sharply with web-based mega-corpora such as GloWbE (Davies & Fuchs 2015), and recent work has shown that differences in compilation procedures and register coverage between ICE and GloWbE have substantive consequences for analyses of World Englishes (Loureiro-Porto 2017). At the same time, research on sociolinguistic variation increasingly relies on corpus- based methods that must balance representativeness, comparability and scale (Botha & Bernaisch 2025). The scheme of the original ICE text categories, partly modelled on late-20th-century British communicative situations, also shows its age: they do not straightforwardly capture today's digital ecologies in many Inner, Outer and Expanding Circle contexts (Kirk & Nelson 2018). Finally, many of the original categories may be challenging to collect, are not relevant in local contexts, and are very small–arguably too small for reliable analysis.To address these challenges, this paper proposes a revised and simplified "ICE 2.0"design for text categorisation that addresses the workshop's focus on compiling, computationally handling and using ICE corpora in the age of AI. The proposed design omits smaller and outdated categories, advocating a larger, functionally defined scheme (see appendix). ICE corpora following this new design should comprise three million instead of one million words to improve coverage of low-frequency phenomena and enable robust subregister analyses.

Crucially, I suggest that in the age of'AI'(interpreted here, as in the workshop call, to relate to Large Language Models), we should not see AI-assisted or AI-generated texts as 'contamination' to be excluded, but as part of the evolving communicative repertoire. Rather than trying to keep ICE 'AI-free', we should embrace the new realities: collect metadata that distinguishes human-authored, AI-assisted and AI-generated texts, treating them as empirical data in their own right, and accept that text categories such as newspaper writing will often be influenced by AI usage and that this is increasingly becoming a 'feature' of language use rather than a'bug'.

Finally, I would like to outline a funding vision for a DFG Research Unit. This consortium would (1) implement and validate the two-layer register design across several newly and simultaneously compiled ICE components, (2) enhance the study of variation in spoken language in World Englishes through its focus on new ICE 2.0 corpora compiled at the same time, and (3) allow for short-term diachronic comparisons with ICE 1.0 corpora.

With these proposals, this paper provides a realistic roadmap for keeping ICE central to research on World Englishes in the age of AI, in line with the workshop's overarching question of what role ICE corpora can and should play 35 years after their conception.

References

Botha, W. & T. Bernaisch. 2025. World Englishes and sociolinguistic variation. World Englishes 44(1–2), 2–11.

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

Greenbaum, S. 1996. Comparing English worldwide: The International Corpus of English. In S. Greenbaum (ed.), Comparing English worldwide: The International Corpus of English, 3–13.

Oxford: Clarendon Press.

Gut, U. & R. Fuchs. 2017. Exploring speaker fluency with phonologically annotated ICE corpora. World Englishes 36(3), 387–403.

Kallen, J. L. & J. M. Kirk. 2012. SPICE-Ireland: A User's Guide. Belfast: Cló Ollscoil na Banríona.

Kirk, J. & G. Nelson. 2018. The International Corpus of English project: A progress report. World Englishes 37(4), 697–716.

Loureiro-Porto, L. 2017. ICE vs GloWbE: Big data and corpus compilation. World Englishes 36(3), 448–470.

Wunder, E.-M., H. Voormann & U. Gut. 2010. The ICE Nigeria corpus project: Creating an open, rich and accurate corpus. ICAME Journal 34, 78–88.

Note: ChatGPT 5.1 and Gemini 3 Pro were used while revising this abstract,

Workshop : Data Management, Corpora and AI Room E314 (afternoon session)
• 14:00–17:30

Conveners: Sabine Bartsch (TU Darmstadt) • Ilka Mindt (University of Paderborn)

This session examines the infrastructure, curation, and long-term management of corpus data in an era when AI tools both depend on and transform data practices. Contact: sabine.bartsch@tu-darmstadt.de

Fifty years of data management at the Oxford Text Archive WSP

Martin Wynne (University of Oxford, UK)

Corpus linguistics has a long tradition of empirical studies and has been at the forefront of implementing digital technology in research and teaching. The most recent empirical and digital turn raises a number of methodological questions concerning the suitability, representativeness, access, storage and archiving of our data. Furthermore, there are concomitant questions concerning the research processes and expressivity of our findings. Related to this are questions concerning our own accountability in the research process including transparency and reproducibility of the research process as well as data accessibility leading to questions of research data management and the role of data publication. And lastly, there are questions surrounding our agreement on the accepted methodology within our research community as well as the development of the required data literacy among researchers who are, more often than not, simultaneously teachers passing on methodological competencies and values within the community. These questions have become even more acute in the context of growing corpus sizes often collected from less well curated sources and whose sheer size necessitates automated and quantitative ap-

proaches for both annotation and analysis that offer a wealth of new insights, but at the same time pose a challenge for comprehensive analysis and close inspection of language samples. The call for linguistics to critically engage with handling a wider spectrum of ever larger corpora while ensuring transparency and reproducibility of research processes and findings is all the more pressing in the context of large language models and the potential and challenges of artificial intelligence applications in corpus linguistics.
In order to further the discussion of these issues in the research community, we propose a half-day workshop that brings together people working in the areas of corpus analysis, corpus compilation, data collection, data management, archive hosting etc. We would like to turn this workshop into a forum for discussion and exchange on desiderates, best-practices, accessibility options and exchange ideas on how data management, methodology and AI use may shape the future.

Automatic Speech Recognition: 3+1 Stairways to Heaven

WSP

Christoph Draxler (LMU Munich, Germany)

Meta and OpenAI give away their AI-based speech models for free: users now have easy access to industry-power spoken language processing, in particular to automatic speech recognition (ASR). However, obstacles remain: a) ASR is often provided by commercial third parties and thus expensive or dangerous with respect to privcay, b) commercial ASR does not necessarily meet academic requirements, c) humanities scholars often do not have the resources to install and maintain ASR solutions, and d) ASR is often not well-integrated into linguistic or humanities workflows.
In this workshop, I will compare current state-of-the-art ASR solutions, present a tool to manage transcription workflows, and outline potential legal issues – all with a focus on use in academia.
The main dimensions of the comparison are the tradeoff between complexity and flexibility, and privacy. The ASR solutions fall into one of three categories on both dimensions: full, pre- and zero configuration for complexity, and strictly local, academic and commercial processing for privacy. I will present standalone tools like aTrain or MacWhisper, the pipeline web services of the Bavarian Archive for Speech Signals, and the CLARIN TranscriptionPortal.
To integrate ASR into transcription workflows, I suggest to use an administration tool such as the Octra Backend. It was designed with a strong focus on privacy, with a clear separation of projects and user roles with specific privileges, and a number of built-in tools or access to optional external speech processing services.
Finally, AI-based processing entails important legal issues: Who may process what data for what purpose, how can privacy be ensured, what type of informed consent is needed from the collaborators – to name some of the most pressing questions.

Corpus Management and Annotation for Large Multimodal Corpora

WSP

Peter Uhrig (FAU Erlangen-Nürnberg, Germany)

Multimodal corpora typically consist at the very least of audio-visual recordings and corresponding transcripts. In addition, a wide range of metadata and annotation can be collected, which results in very heterogeneous datatypes in various formats. In this presentation, I will present some of the challenges around such datasets, in particular with regard to alignment, word-indexed annotation such as PoS and time-indexed annotations such as go-speech gesture. Furthermore, questions relevant to the indexing and retrieval of relevant instances from such rich multimodal corpora will be discussed.

Integrating insights from AI/LLMs into English-Corpora.org

WSP

Mark Davies (Brigham Young University, US)

I have recently carried out seven detailed studies– dealing with word frequency, phrase frequency, collocates, comparing words (via collocates), genre-based variation, historical variation, and dialectal variation (with more than 100 different comparisons in all) – which discuss how well the predictions of two LLMs (ChatGPT and Gemini) compare with the actual data from large, well-known, publicly-accessible, online corpora. The main conclusion is that LLMs are rather poor at generating results (such as word or phrase frequency data or collocates), but they are very good at categorizing and analyzing corpus data.
Based on the fact that AI / LLMs are so good at categorizing and explaining linguistic data, I have now integrated these tools directly into the interface for all of the corpora from English-Corpora.org. With just one click, the corpus can send collocates, frequency patterns, phrase lists, or concordance lines to any of nine different LLMs— which will almost instantly group, explain, and interpret the data. For example, for the first time, collocates can be grouped semantically, and not just by part of speech or function. These AI-powered insights can even relate to syntactic variation and change, such as the use of the passive in different genres, the placement of negation with have over time, or the use of the "like construction" in different dialects). These AI-powered insights (which are clearly marked as such) appear in the interface, alongside the original corpus results.
In addition, users can have the AI generate word and phrase lists for any topic (even abstract topics, like new inventions in the 1600s, or professions in the 1800s that were done mainly by women, or words or phrases related to native foods in Africa, or LGBTQ+ rights in the 2010s). The LLM will generate a list of 150-200 words and phrases, and then each of these is compared to the corpus (EEBO, COHA, BNC, COCA, GloWbE, TV, NOW, etc) and the results will appear with their corpus frequency (including the frequency in each section of the corpus – genre, decade or year, or country)– all within just a few seconds.

There is additional functionality as well, including the ability to see the AI insights and explanations and categorization in any one of 30 different languages, as well as having the AI responses targeted to the level of the user (such as beginning learner, non-native teacher, professional translator, or advanced researcher). Users can also annotate and save their AI queries, and then retrieve them for later use.

RHINE

MOSELLE

Wednesday
27 May 2026

Wednesday 27 May 2026

9:00–9:30am • Welcome & Opening • E011

9:30–10:30am • Plenary 1 • E011

See Plenary Speakers section

Social Media and Online Communication [1] E113 • 11:00–12:30

Pakistani English on Facebook, Instagram, and Twitter/X

FULL PAPER

Muhammad Shakir (University of Münster, Germany)

The overall aim of the paper is to uncover various ways in which English is used by Pakistani internet users on three social media platforms: Facebook, Instagram, and Twitter/X. I conduct three small studies (one qualitative and two quantitative) for each of the social media platform. For Facebook, the study is qualitative in nature and aims to highlight some use cases where English is employed by Pakistani Facebook users. I present an example of a complain post of a delivery boy which shows mesolectal features of Pakistani English (e.g. lack of subject verb agreement) along with typical CMC features (e.g. no punctuations). I also discuss the use of English in a family visa group where a female participant posts about a very personal issue of marriage and divorce in English. Lastly, I highlight how arrange marriage proposal profiles are written mostly in English, even though all other face to face communication takes place in local languages.

For Instagram, the aim is to study how English is used by Pakistani film/drama celebrities. I collect 100 or so post captions from 72 (36 male + 36 female) celebrity accounts. These are mostly in English but local languages also occur. I use Google Gemini Pro 2.5 to classify the posts in: ads, follower, engagement, hashtags, personal life, professional life, politics/social, quotations, and wishes. The results show that Pakistani celebrities use English for posts belonging to ads, personal life, and professional life. Urdu and mixed language posts also occur, but they are only a small portion of their online content. This indicates that Pakistani celebrities embrace English as the dominant language on the internet while communicating with their fans within Pakistan as well as with those living abroad, including their international fans.

For Twitter/X, I aim to find main discourses and themes prevalent in the English posts of users. Twitter (despite intermittent bans) has remained a very popular platform among

Pakistani youngsters in the last decade or so. For this purpose, I use the data collected by Shakir and Deuber (2023) in the SAOnE corpus to conduct a Lexical Multidimensional Analysis (LMDA) developed by Sardinha and Fitzsimmons-Doolan (2025). I get keywords by comparing the Pakistani Twitter data with similar data from the other five varieties of the corpus. After removing function words, I apply PCA to get co-occurring groups of content words. The LMDA shows that politics and related issues are the main theme of Pakistani English tweets. Pakistani netizens on Twitter appear to be either active members of a political party or are their sympathisers. The then (i.e. in 2021) ruling party Pakistan Tehreek e Insaaf (PTI) as well as other main political parties try to spread their narrative on Twitter, e.g. alleging corruption charges against political rivals.
Overall, the results demonstrate that Pakistanis are more likely to use English on these social media platforms, even for locally oriented or personal topics (e.g. divorce/marriage, politics) which would preferably be discussed in Urdu or other regional languages in face to face conversations.

References

Sardinha, T. B., & Fitzsimmons-Doolan, S. (2025). Lexical Multidimensional Analysis: Identifying Discourses and Ideologies (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781009335683

Shakir, M., & Deuber, D. (2023). Compiling a corpus of South Asian online Englishes: A report, some reflections and a pilot study. ICAME Journal, 47(1), 119–139. https://doi.org/10.2478/icame-2023-0007

Linguistic Profiling of Ransomware Actors: Corpus Approaches to Online Communication

FULL PAPER

Christa Schneider (University of Bern, Switzerland)

The increasing availability of large-scale digital communication data has enabled corpus linguists to explore previously inaccessible domains of language use. This paper presents a novel corpus-driven study of ransomware communication, based on a unique dataset of approximately 400,000 messages exchanged between ransomware groups and their victims, or commercial platforms in the Darknet. The corpus, provided legally by Swiss Federal Special Forces for Cyber Crime, represents a unique systematic collection that allows for an unprecedented linguistic exploration of cybercriminal discourse.
The primary objective of this study is to develop a computational-linguistic pipeline for author profiling in the ransomware domain. The pipeline aims to identify recurring stylistic and lexical features that may help define whether individual actors participate in multiple ransomware groups. Drawing on established approaches from corpus linguistics and stylometry, the analysis combines keyword and collocation profiling, register-sensitive feature extraction, and embedding-based clustering. These methods are integrated into an

interpretable workflow designed for forensic and linguistic transparency.
This study will concentrate on a sub-corpus that is of particular interest. It consists of ransom advertisements in English, in which perpetrators seek to sell stolen data to the highest bidder. Preliminary observations reveal that these texts share striking discursive and pragmatic similarities with legitimate commercial advertisements. Both rely on evaluative language, trust-building rhetoric, and formulaic phraseology designed to convey reliability and exclusivity. The quantitative comparison of the commercial strategies allows us to situate ransomware discourse within a broader linguistic economy of persuasion and transaction. The paper will address the following research questions:
1. What linguistic features characterise ransomware communications and distinguish them from legitimate online transactions?
2. To what extent can stylistic and lexical profiling indicate cross-group authorship?
3. How can corpus-based linguistic profiling support digital forensic investigations without compromising ethical and legal standards?
Expected results include an empirically grounded typology of ransomware discourse and a replicable methodological framework for profiling anonymous online authors. Beyond its forensic relevance, this study contributes to corpus linguistic research by extending established methods to a socially and ethically complex communicative context. It also illustrates how AI-assisted corpus analysis - particularly feature extraction and clustering models - can be adapted for transparency and interpretability in applied settings. By situating ransomware discourse within the broader field of digital corpus research, the study exemplifies the conference theme "A Confluence of Corpus Research in the Age of AI": it demonstrates how computational, linguistic, and societal perspectives can converge to illuminate new domains of human communication.

References

Belahcen, M., Camara, D., & Genoe, R. (2022, April). Stylometry: Authorship Identification in Web Forums using Natural Language Processing. Preprint. https://doi.org/10.13140/RG.2.2.24015.76966

Biber, D., Egbert, J., & Davies, M. (2021). Register Variation Online. Cambridge University Press.

Brezina, V. (2018). Statistics in Corpus Linguistics: A Practical Guide. Cambridge University Press.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology, 60(3), 538–556.

Leemann, A., Perkins, R., Buker, G. S., & Foulkes, P. (2025). An introduction to forensic phonetics and forensic linguistics (1st ed.). Routledge

Adding Network Information to Social Media Corpora

FULL PAPER

Mikko Laitinen, Masoud Fatemi & Mehrdad Salimi (University of Eastern Finland, Finland)

Large-scale social media data have enabled corpus linguists to explore new research ques-

tions, for instance in areal variation (Eisenstein 2017), spatiotemporal variation (Grieve et al. 2016, 2018), and socio-demographic variation (Gonzales 2024). We argue that despite these advances, the potential of social media remains underutilized in corpus-based studies, as most studies overlook its role in community formation and network building. This epistemological stance reduces social media to a large-scale digital corpus analyzed through traditional methods.

We introduce an algorithmic method for constructing social networks based on interactional data available in many social media applications. The presentation first demonstrates that network information is accessible, though not always directly, and should play a more prominent role in corpus-based studies. A growing body of research across disciplines has highlighted the fundamental importance of social networks to human life, including language use (Dunbar 2020; Waldinger & Schultz 2023). It is therefore both timely and necessary to integrate social network perspectives more fully into corpus linguistics.

While computational linguistics has begun to examine language use in digital networks (Del Tredici & Fernandez 2018; Würschinger 2021; Zhu & Jurgens 2021), corpus-based studies remain scarce. This raises a key question: why has the integration of ego-network information into social media research been so limited? One explanation lies in methodological challenges. The field currently lacks common approaches for constructing networks from interactional data, and structuring such data in meaningful ways remains difficult (Eisenstein et al. 2014: 11). This presentation addresses that methodological gap. We introduce a methodology for quantifying personal networks using interactional metadata. Our data are drawn from one social media application, but the methods are applicable to any environment where interactions form a directed graph, i.e., where the direction of information flow is observable. The methodology has been developed within a large, externally funded digital infrastructure project, bringing together sociolinguists and computer scientists. We aim to answer two research questions:

1. Informed by recent advances in network science, to what extent can we quantify the structure of ego networks?
2. What types of research potential does the combination of social networks and large-scale social media data afford?

To illustrate this potential, we investigate lexical innovation in English. Prior studies have shown that new lexis emerges from urban centers that act as “hubs of lexical innovation” (Grieve 2018: 309), but we still do not know who the innovators are or in what types of social settings they operate. We use three very large social media datasets containing 3.42 billion words of text from 2020 to mid-2022, representing American, Australian, and British English. The empirical analysis traces emerging lexis (e.g., rizz, oomfies, watchalong, girlies) in networks, showing how areal variation in lexical emergence can be enriched by social information. The presentation offers insights for corpus linguists on how social media data can be accompanied by interaction-based social information. Theoretically, it adds a novel angle to the study of lexical innovation by providing the social context required to understand diffusion

References

Del Tredici, M., & Fernández, R. (2018). The Road to Success: Assessing the Fate of Linguistic Innovations in Online Communities. In E. M. Bender, L. Derczynski, & P. Isabelle (Eds.), Proceedings of the 27th International Conference on Computational Linguistics (pp. 1591–1603). Association for Computational Linguistics. https://aclanthology.org/C18-1135

Dunbar, R. I. M. (2020). Structure and function in human and primate social networks: Implications for diffusion, network stability and health. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 476(2240), 20200446. https://doi.org/10.1098/rspa.2020.0446

Eisenstein, J. (2017). Identifying regional dialects in on-line social media. In C. Boberg, J. Nerbonne, D. Watt (Eds.). The Handbook of Dialectology (pp. 368–383). Wiley. https://doi.org/10.1002/9781118827628.ch21

Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. P. (2014). Diffusion of lexical change in social media. PLOS ONE, 9(11), e113114. https://doi.org/10.1371/journal.pone.0113114

Gonzales, W. D. W. (2024). When to (not) split the infinitive: Factors governing patterns of syntactic variation in Twitter-style Philippine English. English Language & Linguistics, 28(2), 305–339. https://doi.org/10.1017/S1360674323000631

Grieve, J., Nini, A., & Guo, D. (2016). Analyzing lexical emergence in Modern American English online. English Language and Linguistics 21:1, 99–127. doi:10.1017/S1360674316000113

Grieve, J., Nini, A., & Guo, D. (2018). Mapping lexical innovation on American social media. Journal of English Linguistics 46:4, 293–319. doi:10.1177/0075424218793191 .

Waldinger, R., & Schulz, M. (2023). The Good Life: Lessons from the World's Longest Scientific Study of Happiness. Simon & Schuster.

Würschinger, Q. (2021). Social networks of lexical innovation. Investigating the social dynamics of diffusion of neologisms on Twitter. Frontiers in Artificial Intelligence, 4. https://doi.org/10.3389/frai.2021.648583 .

Zhu, J. & Jurgens, D. (2021). The structure of online social networks modulates the rate of lexical change. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics June 6–11, 2021. Human Language Technologies, 2201–2218.

Syntactic Complexity in Academic Writing: Exploring the Influence of Discipline and Authors' L1

FULL PAPER

Mohsen Shirazizadeh (Sohar University, Oman) & Arefe Amini Faskhodi (Alzahra University, Tehran, Iran)

Disciplinary variation in academic writing has long attracted scholarly attention, as disciplines encode distinctive epistemologies through linguistic forms and rhetorical choices. Writers across fields employ diverse rhetorical strategies and linguistic resources to construct and legitimize knowledge, highlighting the central role of disciplinary conventions in

shaping academic discourse. While research on disciplinary variation in academic writing has predominantly focused on surface-level textual features such as specialized vocabulary, phraseological patterns, rhetorical organization, and metadiscoursive markers, there is increasing evidence that the influence of disciplinary conventions extends beyond such choices and are also manifested in deeper levels of text quality such as syntactic complexity (Casal et al., 2021). Despite this, our understanding of syntactic complexity in disciplinary writing remains limited. On the one hand, most studies on syntactic complexity have focused on only one part-genre of research papers (e.g., introduction, discussion), offering only a partial view of how syntactic complexity is realized in an article as a unified whole and as a complete representation of disciplinary writing. On the other hand, the influence of authors' first language on syntactic complexity across disciplines remains largely underexplored, leaving unresolved questions as to how authors' L1 background interacts with disciplinary conventions in shaping the syntactic architecture of academic texts.

To address these gaps, the present study analyzes published research articles from two disciplines, Applied Linguistics and Biology, representing the soft-applied and hard-pure sciences, respectively, written by L1 English and Iranian L2 English authors. The corpus consists of 400 English articles published in international journals, evenly divided across disciplines and author groups. Syntactic complexity was examined using Coh-Metrix 3.0 (McNamara & Graesser, 2012), which generated seven indices capturing key aspects of syntactic sophistication, including two readability measures, left embeddedness, phrase-level modification, two syntactic variety measures, and passive voice density. The data were analyzed using a multivariate analysis of variance (MANOVA) to examine disciplinary differences across both author groups.

Results revealed disciplinary variation in syntactic complexity across the majority of measures, with the two disciplines exhibiting different profiles of complexity across indices. Biology articles exhibited higher complexity in three indices, including passive voice density, phrase-level modification, and L2 readability, reflecting denser and more cognitively demanding structures. In contrast, Applied Linguistics articles showed greater complexity in left embeddedness and both syntactic variety measures, indicating more structurally varied and elaborated clause-level constructions. These patterns were consistent across both L1 groups, suggesting that disciplinary conventions override authors' linguistic background, with writers across L1 groups consistently following established disciplinary patterns in syntactic complexity. The study contributes to understanding cross-disciplinary discourse patterns and offers implications for English for Academic Purposes instruction and cross-disciplinary writing research.

References

Casal, J. E., Lu, X., Qiu, X., Wang, Y., & Zhang, G. (2021). Syntactic complexity across academic research article part-genres: A cross-disciplinary perspective. Journal of English for Academic Purposes, 52, Article 100996. https://doi.org/10.1016/j.jeap.2021.100996

McNamara, D. S. & Graesser, A. C. (2012). Coh-Metrix: An automated tool for theoretical and applied natural language processing. In P. McCarthy & C. Boonthum-Denecke (Eds.), Applied

Natural language processing: Identification, investigation, and resolution (pp. 188-205). IGI Global. https://doi.org/10.4018/978-1-60960-741-8.ch011

Here's when non-alternating examples can be included in alternation research: with the right predictive modeling approach

FULL PAPER

Stefan Th. Gries (UC Santa Barbara & JLU Giessen) & Nina S. Funke (JLU Giessen)

One fruitful area of corpus-linguistic research is alternation research; frequently-studied examples include the dative alternation (*John gave Mary his book* vs. *John gave his book to Mary*), the genitive alternation (*John's book* vs. *the book of John*), particle placement (*John gave back the book* vs. *John gave the book back*), and adjective comparison (*commoner* vs. *more common*). Nowadays, the typical corpus-linguistic study of such alternations involves retrieving examples of the alternants, their annotation for features/predictors that are suspected to influence the choice of alternant, and the analysis of this with some predictive modeling technique – often (mixed-effects) models. Given the emphasis on corpus-based and quantitative analysis, the retrieval of the alternants in the corpus data naturally plays a primary role. A notion quite influential in this context is what in variationist sociolinguistics has been called the idea of "circumscribing the variable context" (Poplack & Tagliamonte 1989:60). Tagliamonte (2012) put it concisely: "Contexts that do not vary but are categorically encoded with one or other variant are not included in the analysis of variation. These are the 'don't count' cases (see Blake 1994). [...] categorical contexts cannot be part of an analysis of variation." This notion has then also influenced retrieval in alternation studies outside of variationist sociolinguistics; e.g.,

- In Grafmiller's (2014) study of the genitive alternation "all tokens involving pronominal possessors were excluded from the data set" because they are "nearly categorical in their preference for the prenominal position";
- Cheung & Zhang's (2016) study of adjective comparison removes "[a]djective types whose analytic-to-synthetic or synthetic-to-analytic ratio is smaller than 0.005";
- a reviewer recently rejected a ms on adjective comparison as "fatally flawed" because it included 1- and 3- syllable adjectives that the reviewer considered "tokens that are not variable in principle."

In addition to the theoretical motivation, this methodological principle is also related to the statistical problem of *complete separation*, the situation when in particular regression modeling approaches (e.g., Varbrul or regular 'glm's) struggle with situations where (combinations of) predictors are only ever attested with one level of the response variable. Both methodologies address the same underlying problem: certain regions of the data space contain no genuine variation and including these regions distorts the analysis. Here,

we demonstrate using simulated data (*n*=300, 2 predictors) and authentic corpus data (from the SAVE corpus, *n*=3535, 6 predictors) regarding the effect of adjective length on comparison that excluding categorical contexts may

- be problematic assumed in how it leads to underestimating the role of very strong predictors (e.g., adjective length, which is (near) categorical for non-disyllabic adjectives);
- not even apply when especially tree-based methods are used to model an alternation. We demonstrate how classification trees using deviance reduction for its splits, deal with non-variable contexts straightforwardly because they
- immediately recognize, and split on, variables leading to (near) categorical prediction;
- continue to use all (combinations of) predictors for all (more) variable contexts.

We propose that this kind of logic be used instead of the traditional discarding of (near) categorical contexts.

References

Blake, Renée. 1994. Resolving the don't count cases in the quantitative analysis of the copula in African American Vernacular English. Paper presented at Stanford University.

Cheung, L., & Zhang, L. 2016. Determinants of the synthetic-analytic variation across english comparatives and superlatives. English Language and Linguistics 20(3). 559-583.

Grafmiller, Jason. 2014. Variation in english genitives across modality and genres. English Language and Linguistics 18(3). 471-496

Poplack Shana & Sali A. Tagliamonte. 1989. There's no tense like the present: Verbal -s inflection in Early Black English. Language Variation and Change 1(1). 47-84.

Tagliamonte, Sali A. 2012. Sociolinguistics as Language Variation and Change. In Sali A. Tagliamonte (ed.), Variationist sociolinguistics: Change, observation, interpretation, 1-24. Hoboken, NJ: Wiley-Blackwell.

Sibilant (de-)voicing in Cameroon English: a study based on a phonemically-annotated corpus

FULL PAPER

Polina Kashkarova & Ulrike Gut (University of Münster)

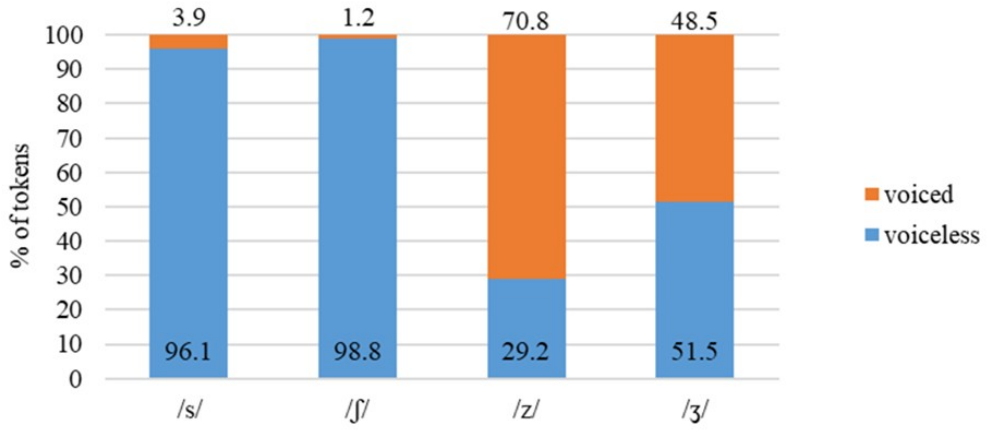
Cameroon English (CamE) has been variously reported to exhibit voicing variation in word-medial sibilants (e.g., Ebot 1999, Simo Bobda 2008, Kouega 2013). For example, [s] for /z/ has been observed intervocalically (e.g., *acqui[s]ition*), [z] for /s/ occurs intervocalically (e.g, *di[z]agree*), after /r/ (e.g., *nur[z]ery*), after /l/ (e.g., *compul[z]ory*) and after other consonants (e.g., *con[z]ume*) as well as in preconsonantal position as in *hu[s]band*. /ʃ/ for /ʒ/ has been reported before *-u* in words such as *ca[ʃ]ual* and before *-ion* in words such as *conclu[ʃ]ion*. These descriptions are, however, based mostly on observation, while the only empirical study to date (Ngefac 2008) analysed a very small set of words and word-list reading data only. Similarly, potential conditioning factors, including phonological environment (Kouega 2013; Simo Bobda 2008), cross-linguistic influence (Ebot 1999: 169), colonial input, teaching traditions, migration-related language contact (Awonusi 1986; Simo Bobda 2003) and CamE-specific phonological processes (Simo Bobda & Chumbow 1999: 43; Simo Bobda 2008: 127), have been proposed in the literature but have not yet been tested empirically in a systematic quantitative analysis.

This corpus study aims to both provide an in-depth quantitative account of the frequency and distribution of (de)voicing of word-medial sibilants in CamE and to investigate factors governing their variable voicing and devoicing. Language-internal factors under analysis include phonetic environment, stress and syllable position. Extralinguistic factors include gender and speech style. The study is based on a corpus of spoken CamE modelled after the spoken part of the ICE corpora (Greenbaum & Nelson 1996). It includes speech of 66 educated CamE speakers in the categories broadcast news, broadcast interviews, broadcast discussions and broadcast speeches. The data were sourced from Cameroonian television programmes on YouTube; the recordings were subjected to automatic speech recognition via WhisperX (Bain et al. 2023) and automatic phone-level forced alignment via the Montreal Forced Aligner (McAuliffe et al. 2025), with subsequent manual correction of the resulting phonemic transcriptions. The dataset comprises 3700 word-medial sibilant tokens. The realisations were determined through auditory-acoustic analysis by trained raters, and Bayesian logistic mixed-effects models were fitted separately for each sibilant.

The results show that voicing of the voiceless sibilants is rare: only 3.9% of /s/ tokens and 1.2% of /ʃ/ tokens were realised as voiced (see Figure 1). By contrast, devoicing of voiced sibilants is substantially more frequent: /z/ was devoiced in 29.2% of cases, and /ʒ/ was the most variable sibilant, with 51.5% voiceless realisations. The modelling results indicate that /s/ voicing is more likely in unstressed syllables, less likely for female speakers and less likely in unscripted speech, while /z/ devoicing is more likely in coda position. Across the

analyses, word-level random effects were large, and many variants were concentrated in a limited set of lexical items. These findings suggest that word-medial sibilant (de)voicing in CamE is lexically patterned rather than the result of a broad phonological rule.

Figure 1. Realization of word-medial sibilants, *N = 3700*



References

Awonusi, V. O. (1986). Regional accents and internal variability in Nigerian English: A historical analysis. *English Studies*, 67, 555–560.

Bain, M., Huh, J., Han, T., & Zisserman, A. (2023). WhisperX: Time-accurate speech transcription of long-form audio. In *Interspeech* 2023 (pp. 4489–4493). International Speech Communication Association. https://doi.org/10.21437/interspeech.2023-78

Ebot, W. (1999). Phonological peculiarities in Cameroon English. *English Studies*, 80(2), 168–179. https://doi.org/10.1080/00138389908599173

Greenbaum, S., & Nelson, G. (1996). The International Corpus of English (ICE) project. *World Englishes*, 15, 3–15. https://doi.org/10.1111/j.1467-971x.1996.tb00088.x

Kouega, J.-P. (2013). RP and the Cameroon English Accent: An Overview. *US-China Foreign Language*, 11(12), 887–900. https://doi.org/10.17265/1539-8080/2013.12.001

McAuliffe, M., Fatchurrahman, M. R., Feiteng, GalaxieT, NTT123, Gulati, A., Coles, A., Kong, C., Veaux, C., Eren, E., Gritskevich, E., Thor, G., Mishra, H., Ogasawara, H., Fruehwald, J., Maria, Potrykus, P., Jung, S., Sereda, T., … 彭震东. (2025). *Montreal Forced Aligner* (Version 3.3.6) [Software]. Zenodo. https://doi.org/10.5281/zenodo.17203348

Ngefac, A. (2008a). *Social differentiation in Cameroon English: Evidence from sociolinguistic fieldwork*. Peter Lang.

Simo Bobda, A. (2003). The formation of regional and national features in African English pronunciation. *English World-Wide*, 24, 17–42.

Simo Bobda, A. (2008). Cameroon English: phonology. In R. Mesthrie (ed.) *Varieties of English 4. Africa, South and Southeast Asia*. Mouton, pp. 115–132. https://doi.org/10.1515/9783110208429.1.115

Simo Bobda, A., & Chumbow, B. S. (1999). The trilateral process in Cameroon English phonology: Underlying representations and phonological processes in non-native Englishes. *English World-Wide*, 20, 35–65. https://doi.org/10.1075/eww.20.1.02sim

Regional patterns of vowel variation in educated West African English: A corpus phonetic study

FULL PAPER

Philipp Meer (University of Münster)

The World Englishes paradigm has had a traditional research focus on the national level. This applies to several well-known models of World Englishes, including Kachru's (1985) Three Circles Model and Schneider's (2007) Dynamic Model. A considerable body of research has also investigated global aspects of English (e.g. Szmrecsanyi & Kortmann 2009; Rutter 2011; Gonçalves et al. 2018; Lawson et al. 2019). However, regional aspects – both at the sub-national as well as cross-national level – have been less considered in empirical research. While older models of World Englishes, such as McArthur (1987) and Görlach (1990), take into account the regional domain and propose the existence of pan-regional varieties in geographical contexts like English-speaking West Africa, systematic research remains sparse. Specifically, while several phonetic and phonological communalities of English as spoken in West Africa have been described (e.g. Awonusi 1986; Simo Bobda 1995, 2003; Gut 2012), the empirical basis for such claims is limited (but see e.g. Brato & Huber 2020; Gut & Meer 2025, 2026; Meer 2026).

The present study aims to address this gap by providing a corpus phonetic investigation into regional patterns of variation in educated English spoken in West Africa (Nigeria, Ghana, and Cameroon) – with a focus on the realization of monophthongs as well as both the sub-national and cross-national levels. To that end, the speech of 96 Nigerian and 39 Ghanaian speakers taken from ICE-Nigeria and ICE-Ghana is analyzed. As an ICE sub-corpus for Cameroon is not available, we compiled a corpus of spoken Cameroon English in line with the design of ICE. The Cameroonian corpus was built using AI-based Automatic Speech Recognition (ASR) and speaker diarization with WhisperX (Bain et al. 2023), followed by manual corrections. A total of 112 Cameroonian speakers were included. Automatic phonetic alignment was performed using FAVE-align (Rosenfelder et al. 2014), which has been shown to perform reliably on New Englishes speech data, approaching human inter-rater agreement in vowel segmentation (Meer 2020). Spot checks confirmed general accuracy. Drawing on Bayesian vowel formant estimation for reliable large-scale acoustic analysis (Meer et al. 2021), F1 and F2 were extracted, and vowel duration was measured. Formant estimation for the Nigerian and Ghanaian data is complete; the Cameroonian vowel formant data is currently being extracted and analyzed. Statistical modeling of vowel formant and vowel duration data employs linear mixed-effects modeling, MANOVA, and random forest analysis. Preliminary results show systematic differences in vowel realization between Northern and Southern Nigerian English and Ghanaian English (see Figure 1). Differences in vowel duration also exist (see Figure 2). With the Cameroonian data currently being analyzed, the study will provide the first comparative, corpus phonetic account of educated English across three West African countries. By identifying both shared and divergent patterns, the study contributes to a more fine-grained understanding of regional differentiation in West Africa and offers new empirical evidence for theorizing the role of the region in the World

Englishes paradigm.

References

Awonusi, V. O. (1986). Regional accents and internal variability in Nigerian English: A historical analysis. English Studies, 67(6), 555–560. https://doi.org/10.1080/00138388608598483

Bain, M., Huh, J., Han, T., & Zisserman, A. (2023, March 1). WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. Retrieved from http://arxiv.org/pdf/2303.00747

Brato, T., & Huber, M. (2012). English in Africa. In R. Hickey (Ed.), Areal features of the anglophone world (pp. 161–186). Berlin: De Gruyter. https://doi.org/10.1515/9783110279429.161

Gonçalves, B., Loureiro-Porto, L., Ramasco, J. J., & Sánchez, D. (2018). Mapping the Americanization of English in space and time. PloS One, 13(5), 1-15. https://doi.org/10.1371/journal.pone.0197741

Görlach, M. (1990). Studies in the history of the English language. Heidelberg: Winter.

Gut, U. (2012). Standards of English in West Africa. In R. Hickey (Ed.), Standards of English: Codified varieties around the world (pp. 213–228). Cambridge: Cambridge University Press.

Gut, U., & Meer, P. (2025). Consonant clusters in Nigerian English. World Englishes. Advance online publication. https://doi.org/10.1111/weng.12731

Gut, U., & Meer, P. (2026). The NURSE-DRESS merger in Nigerian and Ghanaian English: A corpus-based, multifactorial acoustic study. In V. Werner, J. Schlüter, O. Schützler, G. Knappe, L. Sönning, & F. Vetter (Eds.), Variation and change in English: Integrating the study of internal and external factors. Berlin: De Gruyter.

Lawson, E., Stuart-Smith, J., & Rodger, L. (2019). A comparison of acoustic and articulatory parameters for the GOOSE vowel across British Isles Englishes. The Journal of the Acoustical Society of America, 146(6), 4363. https://doi.org/10.1121/1.5139215

McArthur, T. (1987). The English languages? English Today, 3(3), 9. https://doi.org/10.1017/S0266078400013511

Meer, P. (2020). Automatic alignment for New Englishes: Applying state-of-the-art aligners to Trinidadian English. The Journal of the Acoustical Society of America, 147(4), 2283–2294. https://doi.org/10.1121/10.0001069

Meer, P. (2026). (th)-variation in Nigerian English: A corpus phonetic study. In P. Meer & U. Gut (Eds.), English Corpus Phonetics and Phonology: Current approaches and future directions. Berlin: De Gruyter.

Meer, P., Brato, T., & Matute Flores, J. A. (2021). Extending automatic vowel formant extraction to New Englishes: A comparison of different methods. English World-Wide, 42(1), 54-84. https://doi.org/10.1075/eww.00060.mee

Meer, P., & Gut, U. (Eds.) (2026). English Corpus Phonetics and Phonology: Current approaches and future directions. Berlin: De Gruyter.

Rosenfelder, I., Fruehwald, J., Evanini, K., Seyfarth, S., Gorman, K., Prichard, H., & Yuan, J. (2014). FAVE (Forced Alignment and Vowel Extraction) [Computer software]. Retrieved from https://github.com/JoFrhwld/FAVE

Rutter, B. (2011). Acoustic analysis of a sound change in progress: The consonant cluster /stɹ/ in English. Journal of the International Phonetic Association, 41(1), 27–40. https://doi.org/10.1017/S0025100310000307

Schneider, E. W. (2007). Postcolonial English: Varieties around the world. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511618901

Simo Bobda, A. (1995). The Phonologies of Nigerian English and Cameroon English. In A. Bamgbo□e, A. Banjo, & A. Thomas (Eds.), New Englishes: a West African perspective (pp. 248–268). Ibadan, Nigeria: Mosuro

Simo Bobda, A. (2003). The formation of regional and national features in African English pronunciation: An exploration of some non-interference factors. English World-Wide, 24(1), 17–42. https://doi.org/10.1075/eww.24.1.03sim

Szmrecsanyi, B., & Kortmann, B. (2009). Vernacular Universals and Angloversals in a Typological Perspective. In M. Filppula, J. Klemola, & H. Paulasto (Eds.), Vernacular Universals and Language Contacts (pp. 33–56). Routledge.

Rhoticity in Singapore English: Macro-cultural orientations and linguistic variation in speech in the media

FULL PAPER

Cheryl Yeo (Ludwig Maximilian University of Munich, Germany)

The variable realisation of coda /r/, henceforth referred to as /r/, has been accounted for by the traditional phonological distinction of English varieties into rhotic and non-rhotic types (Wells 1982:218–220), and remains cornerstone in sociolinguistic research (e.g. Labov 1966; Blaxter et al. 2019). Over the past four decades, prior studies on Singapore English (SgE) (Tan and Gupta 1992; Poedjosoedarmo 2000; Tan 2012) have examined whether this historically British-influenced variety is now shifting toward a more rhotic model, i.e., American English; media influence is often cited as a catalyst behind this increasing rhoticity, yet to date, there has been no known SgE study scrutinising speech in the media. Bridging this gap, this study investigates speech of media personalities who stand between global culture and local Singaporean culture, and seeks to answer the research questions:

1. What are the observations on /r/ in SgE speakers who are media personalities of varying ages, ethnicities and genders?
2. Which language-internal and language-external variables are significant predictors of the realised /r/ in SgE?
3. Does /r/ usage in the speech of Singaporean media personalities across various ages and ethnicities pattern the same way?
4. How would the findings of this study connect to the wider context of rhotacisation of other Englishes?

This quantitative, corpus-based study adopts a variationist approach to the analysis of /r/ in SgE and takes into account concepts of indexicality (Eckert 2008, 2019), where indexical links between /r/ variants and social meanings are made salient. Examining the speech of media personalities, who position themselves as conduits of globalisation while still appealing to local audiences, addresses a gap in SgE rhoticity research and allows for an investigation of how /r/ usage reflects macro-cultural/global-local orientations (Alsagoff 2010). All speech

data comprise spontaneous, conversational material sourced from publicly accessible media platforms. Data were transcribed and coded auditorily; a subsample of tokens was analysed acoustically using Praat (Boersma and Weenick 2022). In total, 5581 tokens were annotated for linguistic and social variables.

The full sample of 5581 tokens is analysed by means of mixed-effects logistic regression using R (R Core Team 2024). The factors of phonological context, preceding vowel, ethnicity, dominant language (i.e., speaker's first language), orientation, and speech activity are hypothesised to be significant predictors of rhoticity. Additionally, emergent community grammars are observed by running separate models by ethnicity, where a comparison of significant factors and constraint rankings are used to assess the convergence or divergence of these groups' systems.

Quantitative analyses indicate that rhoticity in SgE is present but uneven. Eurasian and Chinese speakers show the highest /r/ realisation rates and clear sensitivity to phonological environment and preceding vowel: consistent with rhotacisation trajectories in other Englishes; whereas Malay and Indian speakers strongly disfavour /r/, suggesting limited participation in this phenomenon. Qualitative analyses reveal style-shifts across utterances: when adopting a localised orientation rather than a globalised one, speakers gravitate toward non-rhoticity. Taken together, rhoticity is a salient but differentiated feature in the speech of Singaporean media personalities, with emergent community grammars varying across ethnicities.

References

Awonusi, V. O. (1986). Regional accents and internal variability in Nigerian English: A historical analysis. English Studies, 67(6), 555–560. https://doi.org/10.1080/00138388608598483

Alsagoff, Lubna. 2010. English in Singapore: Culture, capital and identity in linguistic variation. World Englishes 29(3): 336-348.

Blaxter, Tam, Kate Beeching, Richard Coates, James Murphy, and Emily Robinson. 2019. Each p[□]son does it th[ɛː] way: Rhoticity variation and the community grammar. Language Variation and Change 31(1):91-117.

Boersma, Paul and Weenink, David. 2022. Praat: doing phonetics by computer [Computer software, Version 6.2.23]. <http://www.fon.hum.uva.nl/praat/>; (accessed 11 June 2022).

Eckert, Penelope. 2008. Variation and the indexical field 1. Journal of sociolinguistics, 12(4), 453–476.

Eckert, Penelope. 2019. The limits of meaning: Social indexicality, variation, and the cline of interiority. Language, 95(4), 751–776.

Labov, William. 1966. The effect of social mobility on linguistic behaviour. Sociological Inquiry 36(2):186-203.

Poedjosoedarmo, Gloria. 2000. The media as a model and source of innovation in the development of Singapore Standard English. In Adam Brown, David Deterding and Low Ee Ling, eds. The English language in Singapore: research on pronunciation, 112-120. Singapore Association for Applied Linguistics.

R Core Team. 2024. R: A Language and Environment for Statistical Computing. R Foundation

for Statistical Computing. Vienna, Austria. [Computer software, Version 4.3.3]. <https://www.R-project.org/> (accessed 18 October 2024).

Tan, Chor Hiang and Anthea Fraser Gupta. 1992. Post-vocalic /r/ in Singapore English. York Papers in Linguistics 16:139-152. Tan, Ying Ying. 2012. To r or not to r: Social correlates of /ɹ/ in Singapore English. International Journal of the Sociology of Language 218:1-24.

Wells, John C. 1982. Accents of English: Volume 1. Cambridge University Press.

Gibraltar English in transition: Language attitudes and identity in a World Englishes context

FULL PAPER

Cristina Suarez-Gomez (University of the Balearic Islands, Spain)

This study examines the evolving linguistic ecology of Gibraltar, focusing on the transition from widespread multilingualism in the 20th century to the increasing dominance of English in the 21st. Traditionally, Gibraltarians used English as the institutional language while employing Spanish and Llanito, a vernacular language combining elements of English and Spanish, among others, for informal interaction. Recent research (Author(s) 2026) suggests that English has become the primary language among younger generations in all situational contexts, reflecting a gradual move toward monolingualism. This linguistic transformation, influenced by globalization and geopolitical developments such as Brexit, has reshaped Gibraltar's social and cultural dynamics (Chevasco 2019).

Despite this shift, many Gibraltarians continue to self-identify as multilingual, revealing a paradox between identity and linguistic practice. The central research questions guiding this study are: (i) How are language attitudes toward English, Spanish, and Llanito discursively constructed in Gibraltar's public sphere? (ii) What do these constructions reveal about Gibraltarian identity? We hypothesize that this paradox reflects the ambivalent nature of language attitudes, which mediate between linguistic practices and identity formation, especially in World Englishes (Schneider 2007; Bernaisch&Koch 2016; Buschfeld&Kautzsch 2020).

To address these questions, the study adopts the societal treatment approach to language attitudes (Walsh 2022), which relies on indirect observation rather than self-reported data. A corpus-assisted discourse analysis has been conducted on a dataset of c.13,000 tokens drawn from the GibPress Corpus (Author(s) fc.), using AntConc software. The data were extracted using over 20 keyterms related to the ethnonyms English, Spanish, Llanito, and language use more generally (cf. Graedler 2014). The analysis follows a qualitative, top-down framework, with interpretations cross-checked by additional coders to reduce subjectivity (cf. De Fina 2011).

Preliminary findings indicate that language attitudes in Gibraltar are complex; the main

attitude towards English is one of pride, which is seen to hold strong prestige and symbolize connection to Britain (e.g. "Multi-cultural English-speaking Gibraltar is one that allows access to the entire English-speaking world and that is a significant benefit to Gibraltar" Panorama 2019). Spanish occupies an ambivalent position, valued for its cultural and familial significance yet problematized by political tensions (cf. "[A]s the post-war period developed and restrictions from General Franco bit harder and harder, the devotion and loyalty of the British People of Gibraltar for Queen Elizabeth as the embodiment of the Crown grew deeper and deeper" Panorama 2022). Llanito, though less frequently mentioned, emerges as a potent emblem of local distinctiveness, often invoked in discussions of belonging and authenticity (e.g. "It is very much ours. We celebrate it with our flag as we are 100% llanito" Panorama 2014). Overall, the corpus reveals how media discourse both mirrors and shapes public perceptions of linguistic and cultural identity, highlighting tensions between political alignment and cultural autonomy.

This research contributes to the understanding of Gibraltar English within the World Englishes paradigm by situating language attitudes within their sociopolitical and media contexts. It demonstrates how the negotiation of multilingualism and identity in Gibraltar reflects broader global processes of linguistic change and identity redefinition.

References

Author(s) 2026

Author(s) fc.

Bernaisch, T. & Koch, C. (2016). Attitudes towards Englishes in India. World Englishes, 35(1), 118-132.

Buschfeld, S. & Kautzsch, A. (2020). Introduction. In S. Buschfeld & A. Kautzsch (Eds.), Modelling World Englishes: A Joint Approach to Postcolonial and Non-Postcolonial Varieties (pp. 1-15). Edinburgh University Press.

Chevasco, D. (2019). Contemporary Bilingualism. Llanito and Language Policy in Gibraltar: A Study. University of Cádiz.

De Fina, A. (2011). Researcher and informant roles in narrative interactions: Constructions of belonging and foreignness. Language in Society, 40(1), 27–38.

Graedler, A-L. (2014). Attitudes towards English in Norway: A corpus-based study of attitudinal expressions in newspaper discourse. Multilingua, 33(3–4), 291–312.

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

Walsh, O. (2022). Discourse analysis of print media. In R. Kircher & L. Zipp (Eds.), Research Methods in Language Attitudes (pp. 19-37). Cambridge University Press.

Collocational Development at the Interface of Interlanguage and Nigerian English: A Corpus and AI-Assisted Study within a World Englishes Framework

FULL PAPER

Peter Obukadeta (Brunel University of London Pathway College, United Kingdom)

While collocational competence is widely acknowledged as a core component of L2 proficiency (Henriksen, 2013; Wolter & Yamashita, 2018), it remains comparatively underresearched within Postcolonial Englishes, particularly in African contexts. Much of the existing work on L2 collocations has been conducted in Asian settings and has tended to focus on EFL university learners, leaving African learner populations largely unexplored. This paucity of research means that collocational behaviour at the intersection of interlanguage development and Nigerian English—one of the major varieties within the World Englishes paradigm — remains insufficiently theorised. However, research on Nigerian English has revealed distinctive lexico-semantic and collocational patterns that challenge conventional accuracy-based accounts of learner performance (Okoro, 2013; Obukadeta, 2019). Against this backdrop, the present study reports on a corpus- and AI-assisted investigation of collocational development in the writing of Yoruba-speaking learners of English in Nigeria. Using the half-million-word Nigerian Learner Corpus of English representing four proficiency levels (Obukadeta, 2019), the study employs AntConc to extract verb–noun and adjective–noun collocations and to calculate standard association measures. These collocational patterns are traced developmentally and interpreted through detailed concordance analysis. To clarify the status of recurrent but non-canonical combinations — relative to Nigerian English usage and not the traditional non-native vs native speaker norms — the study draws on two complementary resources: (1) the Nigeria component of the Corpus of Global Web-Based English (GloWbE) as a reference variety for Nigerian English (Davies & Fuchs, 2015), and (2) ChatGPT as an AI-based informant to support qualitative judgements of naturalness, semantic compatibility, and potential localisation, building on emerging work on the use of large language models in corpus linguistics (Uchida, 2024).

The analysis addresses three interrelated questions: (1) how collocational profiles vary across proficiency levels; (2) the extent to which concordance evidence, AI-assisted interpretation, and comparison with GloWbE can distinguish developmental deviations from accepted Nigerian English norms; and (3) what these patterns reveal about collocational development at the intersection of interlanguage and World Englishes.

This study demonstrates how a large, proficiency-tagged Nigerian learner corpus, combined with targeted concordance analysis and comparison with GloWbE, can reveal developmental patterns in verb–noun and adjective–noun collocations that are specific to the interlanguage–World Englishes interface. It further contributes methodologically by showing how AI — in the form of a large language model — can be integrated with corpus evidence to support nuanced judgements of naturalness and localisation, offering a replicable framework for AI-enhanced corpus research in World Englishes.

References

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

Henriksen, B., (2013). Research on L2 learners' collocational competence and development–a progress report. C. Bardel, C. Lindqvist, & B. Laufer (Eds.) L, 2, pp.29-56.

Obukadeta, P. (2019) Collocations in a Learner English Corpus: Analysis of Yoruba-speaking Nigerian English learners' use of collocations [PhD thesis, Kingston University].

Okoro, O. (2013) 'Exploring Collocations in Nigerian English Usage', California Linguistic Notes, 38(1), pp. 84-121

Uchida, S., (2024). Using early LLMs for corpus linguistics: Examining ChatGPT's potential and limitations. Applied Corpus Linguistics, 4(1), p.100089.

Wolter, B. & Yamashita, J., (2018). Word frequency, collocational frequency, L1 congruency, and proficiency in L2 collocational processing: What accounts for L2 performance?. Studies in second language acquisition, 40(2), pp.395-416.

Translingual Nominal Address Systems of South Asian (Heritage) Authors

FULL PAPER

Anke Lensch (Universität zu Köln, Germany)

This paper illustrates how the nominal address systems attested in contemporary English prose fiction written by authors with Indian and Sri Lankan backgrounds display systematic translingual (cf. Canagarajah 2013: 41) effects of different substratum languages, such as Hindi, Malayalam, Sinhala and Tamil. By way of a quantitative and qualitative corpus analysis of a tailor-made corpus compiled out of contemporary novels amounting to more than 15 million tokens, the paper determines systematic parallels and differences in the nominal address systems represented in these works of fiction.

The analysis shows that irrespective of the linguistic background of the individual authors, the use of nominal address terms in all of the novels reflect that they are systematically employed for paying respect to others, be it strangers, acquaintances or relatives. Thus, e.g., Sir and Madam are often used to address acquaintances and more senior colleagues or teachers (cf. Larina & Suryanarayan 2023: 148; Ramachandra 2024: 16), see (1) and (2):

(1) I used to go to school with your sister, sir. acquaintance
(The Seven Moons of Maali Almeida, Karunathilaka)

(2) 'Ratna Madam, maybe you should stay back here.' more senior colleague
(Hot Stage, Nair)

In addition, kinship terminology is often used fictively (cf. Braun 1988: 9) when addressing individuals who are unrelated but more senior in age and/or rank (cf. Ekanayaka 2020: 345; Larina & Suryanarayan 2023: 148; Ramachandra 2024: 16), no matter whether the addressee is a stranger, as in (3), or an acquaintance, as in (4):

(3) ""Hello, sister," the man said carefully in English. "What is your name please?" (The God

of Small Things, Roy) stranger
(4) "How much sugar would you like, auntie? Which kind of sugar would you like?" (Motherland, Vijayaraghavan) acquaintance

The data shows that when English translations of kinship terms are used, see (3) and (4), linguistic differences that could be attributed to the heritage languages of the respective authors largely remain covert. However, these become overt in borrowings of nominal terms of address originating in South Asian languages, see bhai - Hindi 'brother', see (5), and didi - Hindi 'edler sister', see (6):

(5) 'Thanks, bhai,' the tanker driver said. acquaintance
(Hot Stage, Nair)
(6) "They looked at Sai. "Didi…," the woman said. strangers
(The Inheritance of Loss, Desai)

These borrowings are attested in all of the novels under scrutiny and their use correlates with the heritage language(s) of the respective author(s). The paper thus provides more empirical evidence supporting the observation that Indian English and Sri Lankan English language users with bi- and multilingual identities are developing "translingual norms" by mixing English with local conventions to "display their cultural values" (Larina & Suryanarayan 2023: 147). The parallels carved out in the corpus analysis strongly suggest that regarding nominal address systems, Indian English and Sri Lankan English have arrived at the stage of endonormative stabilization (Schneider 2007: 49). At the same time, the systematic differences that correlate with the respective heritage languages of the individual authors point towards the final stage of differentiation (Schneider 2007: 53)

References

Braun, F. (1988). Terms of address. Problems of patterns and usage in various languages and cultures. Mouton de Gruyter.

Canagarajah, S. (2013). Negotiating translingual literacy. Research in the Teaching of English 48. 40–67.

Ekanayaka, T. (2020). Sri Lankan English. In: Bolton, K.; Botha, W. & Kirkpatrick, A. (Eds.): The Handbook of Asian Englishes (pp. 337–353). Wiley.

Larina, T. & Suryanarayan, N. (2023). Address forms in academic discourse in Indian English. In: Baumgarten, N. & Vismans, R (Eds.) It's different with you: Contrastive perspectives on address research (pp. 142–170). John Benjamins.

Ramachandra, I. C. (2024). Informal Address Practices among University Students in Sri Lanka: A Case Study. Sri Lanka Journal of Social Sciences and Humanities 4 (2): 11–21.

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

A Perfect Annotator? Comparing the Performance of Humans and LLMs as Annotators of Perfect Constructions

FULL PAPER

Bethany Dallas, Sofia Carpentieri & Carlos Hartmann (University of Zurich, Switzerland)

As Large Language Models (LLMs) find increasing application in data extraction and annotation, the promise of LLMs to solve any language task with minimal instructions (Brown et al. 2020: 3) is put to the test. Certainly, they are efficient, with accuracy comparable to or even outperforming human annotation (see Haq et al. 2025; Pavlovic and Poesio 2025). However, LLMs' performance varies considerably depending on the task, the implementation strategy, and LLM biases (see Tseng et al. 2025: 9; Vera and Driggers 2025). This paper aims to test the efficiency of LLMs in high-difficulty, form-independent categorisation tasks, specifically the recognition of constructions with perfect meaning in English sentences. The English perfect aspect is used to express actions in the past with current relevance (Michaelis 1993). Although typically associated with the present perfect construction HAVE + participle (see example 1), other constructions also frequently express perfect meaning. Most common is the simple past (Michaelis 1993: 113) as in (2). However, many other alternative constructions are used, such as the after-perfect in (3), or the present continuous as in (4):

(1) I haven't phoned her yet. (ICE-GB- S1A-100)
(2) see story of my life I never learned to shut up right (ICE-SCOT-dem-12)
(3) The wife and children are after going off there the other day (ICE-IRE-S1A-067)
(4) I am running this pub since 1947 and we never had an accident [...]. (ICE-IRE-W2C-017)

Even for human annotators, establishing whether alternative constructions truly express perfect meaning can present a challenge, as the categorisation is often only interpretable in context. While LLMs as annotators are also coming into play in the world of corpus linguistics (e.g. Bleaman and Kommerell 2024; Jubelius 2025), most existing studies focus on the extraction of variables based on form rather than meaning. Our study aims to tackle the form-meaning mapping from the opposite direction and targets the following research questions:

RQ1: How do different prompting strategies (e.g. zero-shot, one-shot, few-shot or variations in task framing) affect an LLM's performance in meaning-based perfect aspect annotation?
RQ2: How does the accuracy and consistency of LLM annotation compare to human performance?

Using iterative and systematic prompt engineering (see Schulhoff et al. 2024) and factoring in insights and limitations of previous studies (e.g. Haq et al. 2025; Pavlovic and Poesio 2025), we will establish the most efficient methodology for extracting sentences with perfect meaning using large-scale commercial LLMs as well as open-source local models. The data extraction will apply established perfect categories (see Michaelis 1993; Miller 2000) to the International Corpus of English (ICE) subcorpora ICE Ireland, ICE Scotland, and ICE Great

Britain, and test manually for precision and recall. Based on previous findings (e.g. Wang et al. 2024; Author 2025), we expect humans to be flexible and fairly reliable annotators with few weaknesses, and the commercial LLMs to perform at or above human level, despite expected bias towards standard English perfect constructions. Further, we anticipate that the smaller, open-source LLMs will perform worse overall but may adapt to non-standard constructions through prompting or fine-tuning.

References

Gut, U. (2020). ICE Scotland Corpus. University of Münster. https://doi.org/10.17879/06968733776

Nelson, G., Wallis, S., & Aarts, B. (2002). The International Corpus of English – Great Britain. Survey of English Usage, University College London. Accessed via ICE Online, University of Zurich. https://www.es.uzh.ch/en/research/corpling.html

Kallen, J. L., & Kirk, J. M. (2008). The International Corpus of English – Ireland. University of Dublin. Accessed via ICE Online, University of Zurich. https://www.es.uzh.ch/en/research/corpling.html

Bleaman, I. L., & Kommerell, R. (2024). A computational approach to detecting the envelope of variation. Linguistics Vanguard, 10(1), 385-395

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. and Agarwal, S. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

Hartmann, C., (2025) Who is the Pronoun Pro? Comparing LLMs and Trained Human Annotators in Pronoun Resolution. ISLE8 Conference, University of Santiago de Compostela, 02.09.2025.

Haq, M. U. U., Rigoni, D., & Sperduti, A. (2025). LLMs as data annotators: How close are we to human performance. arXiv. https://arxiv.org/abs/2504.15022

Jubelius, L. (2025) Coding Corpus Data with LLMs. A Case Study on German Double Superlatives. Workshop "Corpus linguistics 2040: Which data, which methods, which models?", Leibniz-Institut für Deutsche Sprache (IDS), Mannheim, 11.07.2025

Michaelis, L. A. (1993). Toward a grammar of aspect: the case of the English perfect construction. University of California, Berkeley.

Miller, J. (2000). The perfect in spoken and written English. Transactions of the Philological Society, 98(2), 323-352.

Pavlovic, M., & Poesio, M. (2024). The effectiveness of LLMs as annotators: A comparative overview and empirical analysis of direct representation. arXiv. https://arxiv.org/abs/2405.01299

Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., Li, Y., Gupta, A., Han, H., Schulhoff, S. and Dulepet, P.S. (2024). The prompt report: a systematic survey of prompt engineering techniques. arXiv. https://arxiv.org/abs/2406.06608 .

Tseng, Y. M., Chen, W. L., Chen, C. C., & Chen, H. H. (2025). Evaluating Large Language Models as Expert Annotators. arXiv. https://arxiv.org/abs/2508.07827

Vallejo Vera, S., & Driggers, H. (2025). LLMs as annotators: the effect of party cues on labelling decisions by large language models. Humanities and Social Sciences Communications, 12(1), 1-11.

Wang, X., Kim, H., Rahman, S., Mitra, K., & Miao, Z. (2024, May). Human-LLM collaborative

annotation through effective verification of LLM labels. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (pp. 1-21).

LLMs as linguistic annotators: Applying an evaluate-then-scale approach to the English causative alternation

FULL PAPER

Quirin Würschinger & Laura Hahn (LMU Munich, Germany)

Large language models (LLMs) promise scalable corpus annotation, but reliability and reproducibility remain open questions (e.g., Pangakis et al. 2023; Pavlovic & Poesio 2025; Yu 2025). We evaluate how well LLMs classify the English causative alternation using four variables central to the phenomenon: transitivity, causativity, subject role, and subject animacy (e.g., Heidinger & Huyghe 2024; Merlo & Stevenson 2001):
(1) She opened the door. (transitive, causative, agent, animate)
(2) The door opened. (intransitive, anticausative, patient, inanimate)
We investigate how reliably LLMs annotate these variables in authentic corpus data, which variables are most and least challenging, how prompt design affects accuracy, and how LLM performance compares to that of human annotators. Our broader aim is to test the use of LLMs for an evaluate-then-scale approach to corpus-linguistic studies. LLMs enable scalable, reproducible corpus annotation, but it is essential to first conduct detailed evaluations of their performance. Once quality thresholds are met, analyses can be scaled efficiently, which benefits many study designs, especially those targeting rare or highly variable phenomena where manual annotation is infeasible. The data consist of random samples from the Corpus of Historical American English (COHA; Davies 2010) for nine alternating verbs: close, dissolve, dry, fill, freeze, improve, open, split, and wake (180 sentences per verb; total N = 1,620; based on Haspelmath 1993). We annotate the data with OpenAI's o4-mini model, using a detailed, refined prompt that provides linguistic background on the causative alternation and the variables of interest and includes few-shot examples of correct and incorrect classifications, and a structured output schema to ensure consistent annotations. Expert Gold Standard annotations, produced by a linguist with domain expertise in the English causative alternation, serve as the baseline for our evaluations. We evaluate LLM performance with standard classification metrics for each variable and verb (accuracy, precision, recall, F1), and we assess the consistency of the LLM's annotations via repeated classifications. We also collect human baselines from four linguist annotators with different levels of experience (MA, PhD) to compare with the LLM's performance.
Our results show consistently strong performance across all variables. Prompt design matters: the detailed prompt with few-shot examples substantially outperforms a simpler version. Overall, the LLM performs on par with or better than human linguist annotators. Human agreement baselines (Cohen's κ, Fleiss' κ) quantify inter-rater agreement among humans, contextualise LLM–human agreement, and complement our tests of run-to-run consistency for the LLM. We provide detailed evaluations using confusion matrices and error

analyses, which reveal interpretable patterns (e.g., adjectival vs. verbal readings; implied agents in passives). We contribute a reproducible pipeline for LLM-assisted annotation, evidence that an informed prompt, few-shot examples, and a structured output schema materially improve output quality, and a detailed, transparent evaluation framework. The English causative alternation serves as a focused case study for a general evaluate-then-scale approach; once quality thresholds are met, the same approach and pipeline support scalable, unsupervised studies of other linguistic phenomena.

References

Davies, M. (2010). The Corpus of Historical American English (COHA). https://www.english-corpora.org/coha/

Haspelmath, M. (1993). More on the typology of inchoative/causative verb alternations. In B. Comrie & M. Polinsky (Eds.), Causatives and transitivity (pp. 87–120). John Benjamins.

Heidinger, S., & Huyghe, R. (2024). Semantic roles and the causative-anticausative alternation: Evidence from French change-of-state verbs. Linguistics, 62(1), 159–202.

Merlo, P., & Stevenson, S. (2001). Automatic verb classification based on statistical distributions of argument structure. Computational Linguistics, 27(3), 373–408.

Pangakis, N., Wolken, S., & Fasching, N. (2023). Automated annotation with generative AI requires validation. arXiv. https://arxiv.org/abs/2306.00176

Pavlovic, M., & Poesio, M. (2025). The effectiveness of LLMs as annotators: A comparative overview and empirical analysis of direct representation. arXiv. https://arxiv.org/abs/2405.01299

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.

BERT for large-scale data filtering and annotation: The dative and benefactive alternations in the written BNC 2014

FULL PAPER

Bethany Stoddard (University of Bonn, Germany)

The English dative alternation (John gave Mary the book / John gave the book to Mary) and the closely related benefactive alternation (Emma bought her friend a gift / Emma bought a gift for her friend) exemplify grammatical alternation, where different syntactic forms encode the same meaning. The choice between variants is conditioned by multiple probabilistic constraints, including many properties of the recipient and the theme/beneficiary (e.g. animacy, givenness, length) (Bresnan & Ford 2010; Theijissen et al. 2009)
Corpus-based studies of these alternations typically identify relevant data by extracting instances of alternating verbs, then filtering for relevant constructions using either manual annotation, dependency parsing, or POS-based patterns. However, the latter approaches still require substantial manual correction to ensure accuracy. The benefactive alternation in particular is infrequent relative to instances of alternating verbs, meaning larger datasets

and more filtering are required to find sufficient instances for analysis. Furthermore, annotating semantic and discourse properties of arguments is difficult to automate reliably. Thus, traditional semi-automatic approaches to filtering and annotating argument alternations require several rounds of manual annotation/correction, limiting the size of datasets for feasible analyses. To facilitate identification and annotation of argument-structure alternations efficiently and at scale, we can look toward transformer-based large language models such as BERT. This study addresses the following research question: Can BERT can be used to 1) identify instances of alternating structures and 2) annotate properties of the verb arguments?

Recent work has demonstrated the utility of BERT for syntactic construction identification (Scivetti & Schneider 2025) and semantic annotation tasks such as classification of animacy (Tepei & Bloem 2024) and information-status (Hou 2020). One study has used BERT to identify and annotate instances the dative alternation (Liu et al. 2025), where a different model was created for each categorization task. Building on this line of work, the present study introduces two models: 1) a simple categorization model for filtering out irrelevant instances of the target verbs, and 2) a Multi-Task Learning (MTL) model to identify the arguments and annotate properties of each.

The data consisted of 104,085 instances of give and 34,664 instances of buy extracted from the written British National Corpus (BNC) 2014 (Brezina et al. 2021). First, a BERT classifier was fine-tuned on 4,800 instances to label each token as a double-object construction (DOC), prepositional construction (PC), or irrelevant. This model achieved 99% accuracy (F1 macro = 0.97). Next, an MTL model was trained to (1) detect theme and recipient/beneficiary spans and (2) for each argument, categorize seven binary variables (animacy, definiteness, pronominality, givenness, concreteness, number, and locality/person). With just 400 manually annotated instances of training data, the MTL model achieved an overall F1 score of 0.67 across the sixteen tasks. Performance is expected to improve further with more training data. These results demonstrate that BERT-based multi-task learning can substantially reduce manual effort in identifying and annotating relevant corpus data, while maintaining high accuracy. This approach paves the way for large-scale, multifactorial analyses of argument structure and supports broader applications in corpus-based variation research.

References

Bresnan, Joan & Ford, Marilyn. (2010). Predicting Syntax: Processing dative constructions in American and Australian varieties of English. Language, 86.1, (pp. 168-213).

Brezina, V., Hawtin, A. & McEnery, T. (2021). The written British National Corpus 2014–design and comparability. Text & Talk, 41(5-6) (pp. 595-615).

Hou, Y. (2020). Fine-grained information status classification using discourse context-aware BERT. arXiv preprint arXiv:2010.14759.

Liu, Z., Yang, H. & Wulff, S. (2015). Modeling the Dative Alternation in English Early Child Language. Open Mind, 9 (pp. 1066–1097). https://doi.org/10.1162/opmi.a.10 .

Scivetti, W., & Schneider, N. (2025). Construction Identification and Disambiguation Using BERT: A Case Study of NPN. ArXiv, abs/2503.18751.

Tepei, M. & Bloem, J. (2024). Automatic Animacy classification for Romanian nouns. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) (pp. 1825-1831).

Theijssen, D., Halteren, H. van, Fikkers, K., Groothoff, F., Hoof, L. van, Sande, E. van de, Tiems, J., Verhagen, V. & Zande, P. van der. (2009). A regression model for the English benefactive alternation: An efficient, practical, actually usable approach. LOT Occasional Series, 14 (pp. 115-130).

From Annotation to Automation: Mapping Populist Style through Multidimensional Analysis

FULL PAPER

Julia Schilling (University of Bonn, Germany)

In recent years, the rise of populist communication on social media has transformed the linguistic landscape of political discourse. Yet while much research has examined what populists say, less attention has been paid to how they say it, i.e., whether their language constitutes a distinct communicative style. Consequently, this study asks: To what extent does populist discourse exhibit consistent stylistic tendencies across ideational dimensions such as people-centrism, anti-elitism, and popular sovereignty? To address this question, the study combines manual annotation, transformer-based text classification, and corpus-driven multidimensional analysis (MDA) to identify the stylistic correlates of populist discourse in more than one million tweets spanning the complete Twitter timelines of Donald Trump, Bernie Sanders, Barack Obama, and all Democratic and Republican U.S. Senators (2010–2021). The annotated subset of 1,750 tweets was manually coded following Wirth et al. (2019) for three ideational dimensions of populism: people-centrism, anti-elitism, and popular sovereignty. These human labels served to fine-tune a RoBERTa-base transformer model (Liu et al., 2019) in a multi-label setting. Building on earlier applications of transformer models to populism detection (Erhard et al., 2025; Bonikowski et al., 2022), the present study achieves comparable performance (macro-F1 = 0.715 ± 0.013; AUC-ROC = 0.911 ± 0.010) with substantially less data through iterative active learning and nested cross-validation, ensuring efficiency and robustness. Predicted probabilities were then applied to the full corpus to extend the analysis beyond the manually annotated subset. Linguistic style was operationalized using MDATT 2.0, a custom feature-extraction toolkit developed by the author that builds on Biber's (1988) multidimensional framework and Clarke's (2019) adaptation for social-media data. Seventy-three binary lexico-grammatical features above a 5% frequency threshold were analyzed through Multiple Correspondence Analysis (MCA), a statistical technique for uncovering latent stylistic dimensions in categorical data. The MCA yielded four interpretable stylistic axes:
(1) Vague Referential, contrasting indeterminate vs. concrete language;
(2) Interactive Informational, contrasting interpersonal vs. expository discourse;
(3) Promotional Oppositional, opposing self-presentation to criticism; and

(4) Narrative Persuasive, distinguishing descriptive from mobilizing styles.
Non-parametric rank-sum tests and Cliff's δ effect sizes show that populist tweets cluster toward the referential, informational, oppositional, and persuasive poles of stylistic space. This pattern indicates that populist discourse on Twitter is linguistically marked by precision, argumentation, and mobilizing intent, reflecting a style that emphasizes clarity and confrontation over ambiguity or self-promotion. Importantly, both human annotations and transformer-based predictions show the same directional associations with these stylistic dimensions, demonstrating that, although the model was trained on content-based labels, it implicitly captures the same form-based variation. This convergence highlights the potential of explainable AI to uncover interpretable linguistic structure within transformer representations of political discourse. Methodologically, the study demonstrates how integrating transformer-based content modeling with corpus-driven stylistic analysis reveals the communicative architecture of political language. Substantively, it contributes to debates on whether populism constitutes a coherent register of political communication, showing that its distinctiveness in the digital sphere lies not in slogans but in a consistent pattern of referential clarity, oppositional framing, and persuasive force.

References

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

Bonikowski, B., Luo, Y., & Stuhler, O. (2022). Politics as Usual? Measuring Populism, Nationalism, and Authoritarianism in U.S. Presidential Campaigns (1952–2020) with Neural Language Models. Sociological Methods and Research, 51(4), 1721–1787. https://doi.org/10.1177/00491241221122317

Clarke, I. (2019). Functional linguistic variation in twitter trolling. International Journal of Speech, Language and the Law, 26(1), 57–84. https://doi.org/10.1558/ijsll.34803

Erhard, L., Hanke, S., Remer, U., Falenska, A., & Heiberger, R. H. (2025). PopBERT.Detecting Populism and Its Host Ideologies in the German Bundestag. Political Analysis. https://doi.org/10.1017/pan.2024.12

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. https://arxiv.org/abs/1907.11692

Wirth, W., Wettstein, M., Wirz, D., Ernst, N., Büchel, F., Schulz, A., Esser, F., Weber, E., Dalmus, C., Engesser, S., & Manucci, L. (2019). NCCR Democracy Module II: The Appeal of Populist Ideas and Messages. In National Centre of Competence in Research (NCCR). Challenges to Democracy in the 21st Century. https://doi.org/10.5771/2192-4007-2018-3-338

Are machines more confident than humans? A corpus analysis of stance in Human and LLM texts

FULL PAPER

Nicole Katzir & Natalia Levshina (Radboud University, The Netherlands)

Large Language Models (LLMs) suffer from "epistemic miscalibration," i.e., they generate information with higher verbal confidence than their actual internal probabilities warrant (Ghafouri et al. 2024; Steyvers et al. 2025). This leads humans to misinterpret such outputs as reflecting high confidence (ibid). The gap between a model's internal confidence and the confidence perceived by humans poses a potential risk, given the widespread use of chatbots as sources of information and advice, even in high-stakes domains such as medicine and law, where inaccurate advice may have severe consequences.

While previous studies have examined this issue primarily from a computational perspective, the present research adopts a corpus-linguistics approach, focusing on the category of stance in human and LLM outputs, using the Human-AI Parallel English (HAP-E) corpus (Reinhart et al. 2025). We adopt Biber et al.'s (1999) definition of stance, which encompasses both attitudinal and epistemic stance, the latter including the source of information and the degree of the speaker/writer's commitment to it. The presence and distribution of stance expressions can influence how readers assess the credibility of utterances, for instance, by shaping inferences about the writer's degree of commitment (Degen et al. 2019). This study aims to see if and how human and LLM outputs differ with respect to stance.

Data

The HAP-E corpus includes 12,000 human-authored English texts from six types (academic, news, fiction, spoken, blogs, and television and movie scripts), each paired with continuations generated by six different LMs from the GPT-4o and Llama 3 families. The GPT-4o and two of the Llama-based models are instruction-tuned, i.e., they have undergone further training to better align with users' goals.

Method

An extensive list of 35 lexico-grammatical stance-marking features was compiled based on Biber et al. (2004). The features include verb-, adjective-, and noun-controlled complement clauses, modals, and various adverb types (e.g., certainty, imprecision, and degree adverbs). We used NLP methods, such as POS tagging and dependency parsing, to extract these features from the HAP-E corpus. We compare four different sources: human data, instruction-tuned GPT-4o and Llama-based models, and non-instruction-tuned Llama-based models. Previous research suggests that instruction-tuned models generate output that diverges more from human writing (Reinhart et al. 2025), and exhibit a larger gap between models' verbalized confidence and actual accuracy (Leng et al. 2024), raising the question of whether such differences are also reflected in how they mark stance.

A random forest model is used to identify the features that contribute most to distinguishing

each source type (human/Llama-base/Llama-base (instruction tuned)/GPT-4o (instruction tuned)). Subsequent qualitative analysis will focus on the most distinctive linguistic features, to examine their role in the perceived stance of LLM outputs.

Preliminary results

The random forest classifier trained on stance features achieved accuracy above the baseline, indicating that differences in the use of stance features are systematic. Looking at the distribution of features across source types, we find that the instruction-tuned models are less similar to humans, consistent with previous research. Ongoing analyses will provide more detail on how these differences are realized across specific stance categories.

References

Biber, Douglas, Susan Conrad, Randi Reppen, Pat Byrd, Marie Helt, Victoria Clark, Viviana Cortes, Eniko Csomay & Alfredo Urzua. 2004. Representing language use in the university: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus. Princeton, New Jersey: Educational Testing Service.

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

Degen, Judith, Andreas Trotzke, Gregory Scontras, Eva Wittenberg & Noah D. Goodman. 2019. Definitely, maybe: A new experimental paradigm for investigating the pragmatics of evidential devices across languages. Journal of Pragmatics 140. 33–48. https://doi.org/10.1016/j.pragma.2018.11.015.

Ghafouri, Bijean, Shahrad Mohammadzadeh, James Zhou, Pratheeksha Nair, Jacob-Junqi Tian, Hikaru Tsujimura, Mayank Goel, et al. 2024. Epistemic Integrity in Large Language Models. arXiv. https://doi.org/10.48550/ARXIV.2411.06528

Leng, Jixuan, Chengsong Huang, Banghua Zhu & Jiaxin Huang. 2024. Taming Overconfidence in LLMs: Reward Calibration in RLHF. arXiv. https://doi.org/10.48550/ARXIV.2410.09724

Reinhart, Alex, Ben Markey, Michael Laudenbach, Kachatad Pantusen, Ronald Yurko, Gordon Weinberg & David West Brown. 2025. Do LLMs write like humans? Variation in grammatical and rhetorical styles. Proceedings of the National Academy of Sciences 122(8). e2422455122. https://doi.org/10.1073/pnas.2422455122

Steyvers, Mark, Heliodoro Tejeda, Aakriti Kumar, Catarina Belem, Sheer Karny, Xinyue Hu, Lukas W. Mayer & Padhraic Smyth. 2025. What large language models know and what people think they know. Nature Machine Intelligence 7(2). 221–231. https://doi.org/10.1038/s42256-024-00976-7.

Stance and Engagement in Human and AI-Generated Research Paper Introductions: A Cross-Disciplinary Corpus-Based Comparison

FULL PAPER

Greta Maslauskienė (Vilnius university, Lithuania)

Stance and engagement markers have long been recognized as central resources through

which academic writers construct disciplinary identity, signal authorial presence, and interact with their readers (Hyland 2005; Zhou & Hyland 2024). With the rapid emergence of generative AI tools such as ChatGPT, the landscape of academic writing and its study is undergoing profound change. Recent work has begun to compare the use of stance and engagement features in AI-generated and human-authored student writing (e.g., Hyland 2025), revealing both overlaps and divergences in how authorial voice and reader alignment are achieved.

This study takes a step further by examining professionally authored research papers across multiple disciplines, focusing specifically on the Introduction section - a key site for positioning research and establishing credibility. Using a self-compiled specialized corpus of authentic English research article introductions from the humanities and social sciences, along with a corresponding set of AI-generated introductions produced by ChatGPT, the study investigates how stance (e.g., hedges, boosters, attitude markers, self-mentions) and engagement (e.g., reader pronouns, questions, directives) are deployed across human- and AI-produced contexts.

The analysis draws on Hyland's (2005) stance and engagement framework and applies quantitative corpus techniques alongside qualitative discourse analysis to compare linguistic patterns and pragmatic functions. The study aims to determine whether ChatGPT can approximate the rhetorical nuance and disciplinary positioning typical of expert academic writing, and to what extent AI-generated texts reproduce or distort human patterns of interactional metadiscourse. The findings contribute to ongoing debates on the impact of generative AI on academic communication, offering insights into both the affordances and limitations of AI-mediated academic discourse production. By employing ChatGPT-5, this study extends previous ChatGPT-4-based research (Jiang & Hyland 2025a; b) to assess whether newer models show improved rhetorical sensitivity.

References

Hyland, K. (2005). Metadiscourse: Exploring interaction in writing. London: Continuum.

Jiang, F., & Hyland, K. (2025a). Does ChatGPT argue like students? Bundles in argumentative essays. Applied Linguistics, 46(3), 375-391.

Jiang, F. K., & Hyland, K. (2025b). Rhetorical distinctions: Comparing metadiscourse in essays by ChatGPT and students. English for Specific Purposes, 79, 17-29.

Zou, H., & Hyland, K. (2024). Stance in article highlights: The promotion of Covid‐19 research. International Journal of Applied Linguistics, 34(2), 466-483.

Noun-phrase complexity in AI-generated and human-authored student essays

FULL PAPER

Sylvi Rørvik (University of Inland Norway, Norway)

This study investigates the use of noun-phrase (NP) modification in argumentative essays written by Norwegian learners of English in comparison with AI-generated texts. The analytical approach is based on Biber et al's (2011) framework of phrasal complexity, which groups NP modifiers into stages which have been shown to broadly coincide with developmental levels in academic writing (cf. e.g. Staples et al 2016). The corpus comprises human-authored texts written by Norwegian students in Years 10 and 11 from the TRAWL corpus (Dirdal et al 2022), and by first-year university students from the Norwegian component of ICLE corpus (Granger et al 2009), and AI-generated texts on corresponding topics where the large language model (LLM) 'KI-chat' (Sikt Norwegian Agency for Shared Services in Education and Research n.d.) was instructed to replicate the style of the respective age groups included in the human-authored corpus.

Previous research has shown that AI-generated texts differ from those produced by students in various ways, such as greater reliance on a generic style (Herbold et al 2023, Amirjalili et al 2024, Jiang & Hyland 2025b); lack of a main argument (Goulart et al 2024); a preference for simpler syntactic constructions (Jiang & Hyland 2025a); lower frequency of interactional metadiscourse (Jiang & Hyland 2025c); and greater information density (Berber Sardinha 2024). The present study aims to determine whether NP complexity is a useful tool to distinguish between human-authored and AI-generated texts, using texts by Norwegian learners of English as a test case. To this end, the study aims to answer the following research questions:

1. To what extent do human-authored and AI-generated texts on each "developmental level", i.e. Year 10, Year 11, and first year of university contain similar frequencies of complex NPs?
2. To what extent do human-authored and AI-generated texts display a developmental trajectory towards more sophisticated NP modification?
3. To what extent is the use of specific NP modifiers similar on each "developmental level" in human-authored and AI-generated texts? For instance, what characterizes human-authored texts in Year 10, and how similar are these to the corresponding AI-generated texts?

Preliminary results indicate that the human-authored and AI-generated texts included in the study contain similar frequencies of complex NPs at each level, but that human-authored texts display a trajectory towards more sophisticated types of NP modification, while AI-generated texts seem to become less sophisticated in their use of NP modification as the "developmental level" increases. In other words, the university students use slightly more sophisticated types of NP modification than the Year-10 students do, while the corresponding AI-generated texts display the opposite trend. However, the AI-generated texts overall display greater proportions of the more sophisticated NP modifier types. As regards the most frequently used individual modifiers, there is a remarkable degree of

similarity between human-authored and AI-generated texts at each "developmental level". The preliminary conclusion is therefore that, in general, NP complexity is not a very good discriminator between human-authored and AI-generated learner texts.

References

Amirjalili, F., M. Neysani, & A. Nikbakht. 2024. "Exploring the boundaries of authorship: a comparative analysis of AI-generated text and human academic writing in English literature." Frontiers in Education. https://doi.org/10.3389/feduc.2024.1347421

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

Biber, D., B. Gray, & K. Poonpon. 2011. "Should We Use Characteristics of Conversation to Measure Grammatical Complexity in L2 Writing Development." TESOL Quarterly 45(1), 5–35. https://doi.org/10.5054/tq.2011.244483

Dirdal, H., I. K. Hasund, E.-M. Drange, E. Vold, & E.-M. Berg. 2022. "Design and construction of the Tracking Written Learner Language (TRAWL) corpus: A longitudinal and multilingual young learner corpus." Nordic Journal of Language Teaching and Learning 10(2), 115–135. https://doi.org/10.46364/njltl.v10i2.1005

Goulart, L., M. L. Matte, A. Mendoza, & L. Alvarado. 2024. "AI or student writing? Analyzing the situational and linguistic characteristics of undergraduate student writing and AI-generated assignments." Journal of Second Language Writing 66. https://doi.org/10.1016/j.jslw.2024.101160

Granger, S., E. Dagneaux, F. Meunier, & M. Paquot. 2009. International Corpus of Learner English v2. Presses universitaires des Louvain. Herbold, S., A. Hautli-Janisz, U. Heuer, Z. Kikteva, & A. Trautsch. 2023. "A large-scale comparison of human-written versus ChatGPT-generated essays." Scientific Reports 13. https://doi.org/10.1038/s41598-023-45644-9

Jiang, F., & K. Hyland. 2025a. "Metadiscursive nouns in academic argument: ChatGPT vs student practices." Journal of English for Academic Purposes 55. https://doi.org/10.1016/j.jeap.2025.101514

Jiang, F., & K. Hyland. 2025b. "Does ChatGPT Write Like a Student? Engagement Markers in Argumentative Essays." Written Communication 42(3), 463-492.

Jiang, F., & K. Hyland. 2025c. "Rhetorical distinctions: Comparing metadiscourse in essays by ChatGPT and students." English for Specific Purposes 79, 17-29.

Sikt Norwegian Agency for Shared Services in Education and Research. n.d. KI-chat. Available at https://sikt.no/tjenester/sikt-ki/ki-chat . (Accessed September 3, 2025)

Staples, S., J. Egbert, D. Biber, & B. Gray. 2016. "Academic Writing Development at the University Level: Phrasal and Clausal Complexity Across Level of Study, Discipline, and Genre." Written Communication 33(2), 149-183

The dative alternation through the lens of a generative AI model

FULL PAPER

Mohammad Alenezi (Kuwait University, Kuwait) & Tobias Bernaisch (Justus Liebig University Giessen, Germany)

The dative alternation, i. e. the choice between the double-object construction (e. g. John gave Mary a book) and the prepositional dative (e. g. John gave a book to Mary), is one of the most widely researched alternations of the English language. The alternation has been investigated in earlier historical periods (Zehentner 2019; Wolk et al. 2013), but also within the World Englishes paradigm focusing on large sets of regional varieties (Röthlisberger 2018) or theoretical models like linguistic epicentres (Gries & Bernaisch 2016). These mostly multifactorial corpus-based studies have shown that the choice between both alternants is guided – admittedly among other factors – by manifestations of the Easy First principle (MacDonald 2013), whereby contextually more accessible and structurally less complex objects tend to precede contextually less accessible and structurally more complex objects. Given the consistency of these contextual and structural effects, the present paper explores to what degree said effects also manifest themselves in dative alternation choices by generative AI models and sets out to study these research questions:

- Do dative alternation choices by generative AI models align with dative alternation choices by human beings when structural and contextual factors are controlled for?
- In cases where dative alternation choices by generative AI models and humans do not align, which predictors account for these differences in dative alternation choices?

To explore this, we extracted 2,124 examples of the dative alternation from the Kuwaiti English National Corpus (KENC2023; Alenezi forthcoming). KENC2023 consists of 128 million words across three subcorpora: spoken, written, and online. The spoken component (634,254 tokens) includes speech from 564 speakers (467 Kuwaitis, 97 non-Kuwaitis). The written component (127 million tokens) features newspaper writing from the Kuwait Times and Arab Times. The online subcorpus (341,934 words) comprises blogs, novels, and opinion columns by Kuwaitis. Individual dative alternation examples have been extracted and annotated for the following predictors:

- VERB: GIVE vs. SELL vs. SEND
- PATLENGTH/RECLENGTH: number of characters of patient/recipient
- PATPRONOMINALITY/RECPRONOMINALITY: non-pronominal vs. pronominal patient/recipient
- PATSEMANTICS/RECSEMANTICS: patient/recipient semantics with seven levels (abstract vs. animate vs. communication vs. concrete/inanimate vs. human vs. institution vs. perception/emotion)
- PATACCESSIBILITY/RECACCESSIBILITY: discourse accessibility of patient/recipient
- MODE: speech vs. writing

We featured the contextual and structural information from these annotations along with the concrete forms of the direct and indirect objects in prompts for the market leader in AI

chat bots, ChatGPT. We prompted the recent and cost-effective generative AI model GPT-5 mini with the help of OpenAI's API via an R script to document the dative alternation choices of the generative AI model. Our results show that human and generative AI choices tend to align most when predictors produce relatively clear-cut cues for constructional choices (e. g. when length differences between direct and indirect objects are pronounced, resulting in an obvious preference for one construction over the other). Still, differences between human and generative AI choices surface when such clear-cut cues are absent.

References

Alenezi, M. A. N. (forthcoming). English in Kuwait: Development, status and usage. John Benjamins.

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

MacDonald, M. C. (2013). How language production shapes language form and comprehension. Frontiers in Psychology, 4, 1–16.

Röthlisberger, M. (2018). Regional variation in probabilistic grammars: A multifactorial study of the English dative alternation. Doctoral Dissertation. Leuven: KU Leuven.

Wolk, C., Bresnan, J., Rosenbach, A., & Szmrecsanyi, B. (2013). Dative and genitive variability in Late Modern English. Diachronica, 30, 382–419.

Zehentner, E. (2019). Competition in language change: The rise of the English dative alternation. Berlin: De Gruyter Mouton.

Pragmatics E313 • 14:00–16:15

Marking the Obvious: Discourse-pragmatic Uses of *duh*

FULL PAPER

Veronika Raušová (Charles University, Czech Republic)

This study aims to provide quantitative and qualitative analysis of the discourse-pragmatic use of duh in English computer mediated, casual written discourse (CMWD), represented by a corpus of Reddit posts and associated comments. Duh is usually defined as an interjection signalling feigned/actual ignorance or mockingly indicating that other speaker's proposition states something obvious (e.g. OED, n.d.; Merriam-Webster, n.d.). I approach duh as a pragmatic marker (PM) operating in the interpersonal domain of discourse (Brinton, 1996), primarily encoding a speaker's strong epistemic stance. Building on Andersen's (2014) relevance-theoretic account, duh marks a proposition as obviously true, whether it is oriented towards speaker's own proposition or to a proposition of another speaker, as an echoic reaction. The Reddit corpus (April 2024-June 2025; 165+ million tokens) sampled data from 16 subreddits favouring textual interaction and high audience engagement (e.g., r/AskReddit, r/Showerthoughts, r/AITAH), with posts formulated to elicit stance-laden

responses. Although duh is associated with casual speech, Reddit CMWD exhibits conversational properties (Cutler et al., 2022) making it suitable for this study. The corpus contains 1,307 tokens of duh (RF=7.9 ipm), of which 1,094 were classified as PM uses. Tokens were coded for position (clause-initial, -medial, -final, standalone) and orientation (towards speaker's own proposition, proposition of another speaker, or the speaker themselves). Frequent collocates (e.g., well, like, I mean) and variants (e.g., no duh) were also noted.

Duh appears most often clause-finally (N=450; 41.1%) and as a standalone marker outside sentence structure (N=450; 41.1%). Clause-initial (N=184; 16.8%) and clause-medial (N=10; 0.9%) uses were less frequent. Regarding orientation, duh primarily marks the current speaker's own proposition as obviously true (N=737; 67.4%) (ex.1), whether it is a direct answer or a single proposition within wider a co-text. When oriented towards a prior speaker's proposition (N=328; 30%), duh marks the proposition as obvious (ex.2), which can be read as derisive in some contexts but may also index strong alignment (ex.3). Because duh signals strong speaker certainty, it is readily exploited for ironic/humorous effects, proposing something as obviously true when it is not (ex.1).

(1) R1: As a redditor, I would say, "Where do I get friends?" in the first place.
R2: At the friend store, duh. (RC_1kvjwoh)
(2) R1: Zombies would smell horrible.
R2: Well duh, they 're corpses (RC_1htlbcn)
(3) R1: Wagons are cooler than sedans and SUVs. [...]
R2: Duh. Wagons rule. (RC_1kxmlzn)

Finally, self-oriented uses are marginal (N=29; 2.7%), often expressing self-deprecation after receiving an explanation (ex.4).

(4) R1:"Kicked up a notch" means "made more intense."
R2: Oooh ok yes duh. 😬 [...]. (RC_1fiwiz7)

Position and orientation were closely associated, $\chi^2(6)=358$, $p<.001$, Cramer's $V=0.405$ (N=1,094), with 97.6% clause-final tokens oriented to the current speaker's proposition and 64.7% clause-initial tokens oriented to prior speakers' propositions.

These findings attempt to show that duh functions beyond interjectional glosses and contribute to research on discourse-pragmatic markers, interjections, and other procedural cues, crucial in the age of AI/LLMs, where models must capture pragmatic coordination to achieve human-like discourse.

References

Andersen, G. (2014). Relevance. In K. Aijmer & C. Rühlemann (Eds.), Corpus pragmatics: A handbook (pp. 143–168). Cambridge University Press. https://doi.org/10.1017/CBO9781139057493

Brinton, L. J. (1996). Pragmatic markers in English: Grammaticalization and discourse functions. Mouton de Gruyter. https://doi.org/10.1515/9783110907582

Cutler, C., Ahmar, M., & Bahri, S. (2022). Introduction: The oralization of digital written communication. In C. Cutler, M. Ahmar, & S. Bahri (Eds.), Digital orality: Vernacular writing in online spaces (pp. 3–31). Palgrave Macmillan. https://doi.org/10.1007/978-3-031-10433-6_1

Oxford University Press. (n.d.). Duh, int. In Oxford English Dictionary. Retrieved October 10,

2025, from https://doi.org/10.1093/OED/8422081641

Merriam-Webster. (n.d.). Duh. In Merriam-Webster.com dictionary. Retrieved October 15, 2025, from https://www.merriam-webster.com/dictionary/duh

Advice in American English conversation: a corpus pragmatics study

FULL PAPER

Rachele De Felice (The Open University, UK) & Nele Põldvere (Lund University, Sweden)

Advice is an important, yet sensitive, social action in conversation (why don't you google it). Recently, Põldvere et al. (2022) carried out a corpus study of British English (BrE) conversation, which establishes the range of constructions advisers use to express communicative (in)directness to different degrees. It also found that who gives the advice and how are the strongest predictors of advice uptake. This presentation replicates this study on American English (AmE) conversation, using Little LANA, an approximately 5-million-word sample of the forthcoming Lancaster-Northern Arizona Corpus of American Spoken English (LANA-CASE, Hanks et al. 2024). In particular, we ask whether the same advice-giving constructions are used in LANA-CASE as in BrE conversation. We also 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 focuses on socio-demographic variables chosen to closely match those in Põldvere et al. (2022). The size of Little LANA makes it necessary to extract a sample similar in size and scope to the BrE data (which totalled approximately 450,000 words). This was created by only including conversations between family members, partners, or friends, and by relying on the corpus's own descriptions of the main functions of the conversation. Using the latter, we only included conversations described as 'giving advice and instructions'; 'figuring things out'; 'sharing feelings and evaluations and opinions'. From this subset, a random sample was extracted, featuring no more than one text per speaker, a text size of 1,000-10,000 words, and as wide as possible a spread of demographic and situational variables to ensure representativeness. The resulting subcorpus consists of 455,000 words and 200 different speakers.

Instances of advice were extracted using a combination of automated pattern-based and lexical-based extraction and manual triage of results, and were then annotated following the manual in Põldvere et al. (2022). The annotation procedure focuses particularly on construction type (drawing on utterance type, references to the participants involved, modality, negation, etc.), modification, type of response, and whether the advice is solicited or unsolicited.

Based on the advice-giving instances extracted so far, we estimate that the rate of giving advice in AmE and BrE conversation is roughly similar. Ongoing results so far have yielded 580 distinct instances of advice (compared to 1,234 in the BrE data), showing, among other features, a majority of declarative constructions, particularly those placed relatively low on

the scale of communicative directness, but still a higher rate of rejection and resistance compared to the BrE data. Thus, similar to the findings in Põldvere et al. (2022), indirect, minimally imposing advice does not necessarily lead to the best communicative outcomes. These preliminary results also have to date not revealed any meaningful differences in types of advice-giving constructions between the two language varieties. Furthermore, indefinite pronouns, contrary to expectations, do not appear to be used as advice-giving strategies in this corpus; the reasons for this absence are still being investigated, along with the role of other linguistic and demographic variables.

References

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

Does speech processing load increase in listeners when speakers use novel word combinations? A case-study in multimodal corpus pragmatics

FULL PAPER

Christoph Rühlemann (University of Marburg, Germany)

This study investigates whether speakers' use of phraseologically novel word combinations increases recipients' processing load in spontaneous conversation. Building on the assumption that listeners use accumulated phraseological experience to anticipate the lexico-syntactic trajectory and possible completion of a turn-constructional unit (TCU), the study examines whether pupil size increases when a speaker moves beyond attested cumulative word combinations into phraseologically unattested territory. The analysis draws on question and storytelling data from the Freiburg Multimodal Interaction Corpus (FreMIC), which contains eye-tracking and pupillometric data from naturally occurring interaction. TCUs were manually segmented and annotated, and word-level measures of cumulative n-gram frequency were used to identify a phraseological 0-point: the point at which the cumulative word combination reaches only a single attestation in the corpus. Words before this point were classified as PreZero, and words at or beyond it as PostZero. The number of words following the 0-point was used as a measure of the extent to which the speaker continued into phraseologically unattested territory.

Baseline-corrected and z-standardized pupil size was analyzed at the word level using linear mixed-effects models with random intercepts for participant, file, and TCU nested within file. The full model tested whether pupil size differed between PreZero and PostZero phases, whether this effect differed between question and pre-climax story TCUs, and whether it was modulated by the length of the post-zero continuation. Results showed a positive PostZero–PreZero contrast for question TCUs, while the corresponding contrast for story

TCUs was weaker and statistically uncertain. A follow-up analysis restricted to question TCUs examined whether the PostZero effect could instead be explained by response planning and next-speaker selection. This analysis showed a mixed pattern. The PostZero increase was not observed in Answerers, although a pure response-planning account would predict increases for Answerers in both exclusive and inclusive Selection contexts. Instead, the only reliable PostZero increase occurred among Not-Answerers in inclusive Selection questions, with a smaller but statistically uncertain positive trend among Not-Answerers in exclusive Selection questions. Thus, the follow-up analysis did not support the view that the PostZero effect is simply a masked next-speaker selection or response-planning effect. At the same time, the evidence for a fully selection-independent phraseological effect was limited, since the most diagnostic condition – Not-Answerers in exclusive Selection – showed only a weak, non-significant positive trend.
These findings suggest that speakers' extended movement into phraseologically unattested territory may increase recipient processing load, but that this effect is shaped by the interactional environment. The results are most compatible with a combined account: phraseological processing demands and next-speaker selection may jointly affect recipients' cognitive load, especially in inclusive Selection questions where recipients are treated as possible next speakers even if they ultimately do not answer. More broadly, the study demonstrates the potential of combining usage-based phraseological measures with pupillometry in ecologically valid conversational data.

World Englishes Grammar Research E314 • 14:00–16:15

A sententialist account? Not if I can help it!

FULL PAPER

Yolanda Fernández-Pena (Universidade de Vigo, Spain), Jong-Bok Kim (Kyung Hee University, Korea) & Javier Pérez-Guerra (Universidade de Vigo, Spain)

This study examines not-if constructions, in italics in (1)–(2). Previous studies have explored negative constructions (Cappelle, 2020, 2021; Schmid, 2013) and, more specifically, negative fragments (Kim, 2024; Weir, 2020), but there has been no detailed investigation of these not-if fragmentary constructions.
(1) You know they're gonna detain you. Not if they can't catch me. [COCA: 2014 TV:State of Affairs]
(2) she wasn't going to tell him about this, not if she could help it. [BNC: 1992 FictMys22]
Formally, not-if expressions are not prototypical clauses/sentences but fragments. In fact, they are instances of so-called 'insubordinate' CP structures (Evans, 2007) since they are introduced by complementiser (subordinating conjunction) if and lack main clauses. Additionally, the complementiser is preceded by the mandatory negative item not.
As regards their meaning, on the one hand, not-if constructions can occur after affirmative

sentences, as in (1), where the former rejects the positive statement expressed by the latter by conditioning it on compliance with the proposition conveyed by the insubordinate clauses (i.e., '[They are] Not [gonna detain me] if they can't catch me'). On the other hand, not-if constructions can follow negative sentences (2), where the insubordinate clauses do not condition the preceding negative statements but reinforce their negative polarity. In such cases, the interpretation of not-if clauses is closer to the contingency reading of examples like If possible, you should test all moving parts 'in cases/circumstances when/where' (Quirk et al., 1985: 1086). Along these lines, example (2) can be interpreted as 'she wasn't going to tell him about this–especially not when she could not help it'. Consequently, the interpretation of not-if constructions may differ from that of prototypical conditional if-constructions. This fact challenges deletion-based analyses that rely on reconstruction from (complete) conditional sentences but opens avenues for a constructionalist approach to these constructions, given their specific non-compositional pairing of form and meaning (Goldberg, 2006).

To investigate this perspective, we have analysed the usage, form, meaning and function of not-if fragments in contemporary English with corpora. Our data were retrieved from The Corpus of Contemporary American English (COCA) and the two releases of The British National Corpus: BNC1994 and BNC2014. Approximately 5,000 instances of not-if constructions were obtained. The data were categorised for several linguistic factors, including the position of the not-if fragment, scope of the negation, syntactic/semantic identity with the antecedent, and semantic resolution.

The findings show that the not-if construction is more frequent in American English than in British English, especially in speech-related text types or genres, which is consistent with its interpersonal nature. This is further evidenced by the proportion of fragments with first- and second-person pronominal subjects (54% on average) and the occurrence of disjunctive dependents (certainly, especially) before the not-if introducer (e.g. [She wasn't getting out of this place,] Certainly not if she didn't play ball; COCA: 2007 FIC:Analog). Finally, regarding the semantics, a substantial number of not-if fragments that follow negative sentences do not accept 'sentential' deletion-based reconstruction, which underscores their constructionalist status.

References

BNC Consortium (2007). British National Corpus: XML edition. Oxford Text Archive (accessed June 2024).

Brezina, V., Hawtin, A., & McEnery, T. (2021). The Written British National Corpus 2014. Design and comparability. Text & Talk, 41(5–6), 595–615.

Cappelle, B. (2020). Not on my watch and similar not-fragments: Stored forms with pragmatic content. Acta Linguistica Hafniensia, 52(2), 217–239.

Cappelle, B. (2021). Not-fragments and negative expansion. Constructions and Frames, 13(1), 55–81.

Davies, M. (2008–). The Corpus of Contemporary American English (COCA). www.englishcorpora.org/coca/ (accessed: June 2024).

Evans, N. (2007). Insubordination and its uses. In I. Nicolaeva (Ed.), Finiteness: Theoretical and empirical foundations (pp. 366–431). Oxford University Press.

Goldberg, A. E. (2006). Constructions at work: The nature of generalization in language. Oxford University Press.

Kaltenböck, G. (2016). On the grammatical status of insubordinate if-clauses. In G. Kaltenböck, E. Keizer, & A. Lohmann (Eds.), Outside the clause: Form and function of extra-clausal constituents (pp. 341–378). John Benjamins.

Kim, J. (2024). (Negated) fragment answers in English: A discourse-oriented and construction-based perspective. English Language and Linguistics, 28(3), 553–588.

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.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). A comprehensive grammar of the English language. Longman.

Schmid, H. (2013). Is usage more than usage after all? The case of English not that. Linguistics, 51(1), 75–116.

Weir, A. (2020). Negative fragment answers. In V. Déprez, & M. T. Espinal (Eds.), The Oxford handbook of negation (pp. 441–457). Oxford University Press.

The Effect of Sub-Register in the Rise and Productivity of Complex Prenominal Modifiers

FULL PAPER

Marcus Callies (University of Bremen, Germany), Turo Vartiainen (University of Helsinki, Finland) & Aatu Liimatta (University of Helsinki, Finland)

The frequency of prenominal modifiers has increased in the recent history of both British and American English. This trend has often been discussed with a view to a more compressed style of writing in registers such as academic writing and journalistic prose, a phenomenon also known as "densification" (Leech et al., 2009). Moreover, recent studies on American English found that premodifiers have also increased in complexity (e.g., Günther, 2019) with types like higher-than-average temperatures and easy-to-read instructions becoming significantly more common in the second half of the twentieth century, especially in magazines and news reporting (Vartiainen et al., 2025; Callies et al. 2026). Both registers are subject to densification, and are also characterised by an expressive style of writing, a feature that has been associated with the use of complex prenominal modifiers (CPMs) previously (e.g., Meibauer, 2007). However, earlier studies have not investigated the rise of CPMs from the perspective of sub-register, even though the importance of sub-registers as mediators of linguistic changes has received particular emphasis (e.g., Biber and Gray, 2013).

In this study, we focus on two specific types of CPMs: 1) a comparative type that consist of a comparative adjective followed by than and either an adjective expressing normality (e.g.,

normal, usual) or a past participle expressing anticipation (e.g., expected, predicted) as in better-than-expected U.S. growth data; and 2) a type that features an infinitival construction preceded by a small set of so-called tough-predicates like easy and hard as in easy-to-clean vinyl sole. Our aim is twofold: we examine the precise effect of sub-registers in the rise of the two types of CPMs, and also how this may be linked to their general productivity from a constructional point of view. We use data from the Corpus of Contemporary American English (COCA) which includes information on sub-register for the news register in terms of eight thematic sections (Misc, News_Intl, News_Natl, News_Local, Money, Life, Sports, Editorial) as well as the thematic domains of the different popular magazines included in the corpus. For magazines, we also explore the TIME Magazine Corpus that consists of articles from the TIME magazine published between 1923-2006 with the corpus metadata specifying the thematic section of the magazine in which the respective text sample occurs (e.g., World, Nation, Sport, Economy & Business).

Our findings suggest that at least some functions associated with the two types of CPMs examined may be connected to specific sub-registers. CPMs of the comparative type appear especially useful when reporting economic trends and developments in the business and money sections of newspapers and magazines, while those of the easy/hard-to-V type frequently occur in magazines to describe the qualities of particular products or services in reviews. We thus conclude that the substantial increase in the general use and productivity of CPMs in the late twentieth century can plausibly be connected to their usefulness in certain sub-registers of magazines and news reporting.

References

Biber, D. & Gray, B. (2013). Being specific about historical change: The influence of sub-register. Journal of English Linguistics, 41(2), 104–134.

Callies, M., Vartiainen, T. & Liimatta, A. (2026). A diachronic register-approach to complex prenominal modifiers. In N. Yáñez-Bouza, D. González-Álvarez & E. Rama-Martínez (eds.), Register and Discourse through the Lens of Corpus Linguistics. Amsterdam: Benjamins, 73-92.

Günther, C. (2019). A difficult to explain phenomenon: Increasing complexity in the prenominal position. English Language and Linguistics, 23(3), 645–670.

Leech, G., Hundt, M., Mair, C. & Smith, N. (2009). Change in contemporary English: A grammatical study. Cambridge University Press.

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

Vartiainen, T., Callies, M. & Liimatta, A.(2025). The productivity of the Complex Modifier Construction in World Englishes. English Language and Linguistics, 29(2), 389–410.

Waxing poetic, lyrical, eloquent or stronger: Phraseological variation in World Englishes

FULL PAPER

Olli Silvennoinen (Åbo Akademi University, Finland)

The verb wax retains an unusual argument structure, [SUBJ WAX PREDAdj], a remnant of its earlier copular use meaning 'become, grow' (see Petré, 2012). Consider (1):

(1) Yes, all sports people like to wax lyrical about how 'it's not the money that counts' [...] (GloWbE GB)

As (1) shows, the typical sense of wax in this construction is communication, and the adjective in the predicative slot expresses the manner of speaking; thus, to wax poetic means 'to speak in a(n increasingly) poetic way', often with a slightly derogatory or ironic connotation (OED). According to preliminary observations, the adjectival predicative slot has two roughly synonymous main fillers, lyrical and poetic, the former being favoured in British English and the latter in American English. However, the slot is open to other adjectives as well, and these are often drawn from the semantic field of literature, as in (2), or similar fields, as in (3).

(2) Back home they might wax rhapsodic about a few hours mucking about with bricks and mortar. (GloWbE CA)

(3) But for the past couple of days, Twitterers have united to wax nostalgic on something quite positive: the video games they grew up playing. (GloWbE AU)

This presentation will consider this phraseological variation in the use of wax as a verb in World Englishes, using the Corpus of global web-based English (GloWbE), a 1.9-billion-word corpus representing 20 varieties of English. It asks (i) what kind of collocational patterning the adjectival predicative slot of the wax construction displays, and (ii) how this collocational patterning is in turn patterned in different varieties of English. To do this, all instances of wax followed by an adjective will be collected, and the resulting dataset will be analysed using collostructional analysis and hierarchical agglomerative clustering to see if and how the varieties pattern together (see Mukherjee & Gries, 2009 for a similar approach). Preliminary results confirm that lyrical and poetic are indeed the main collocates of wax in the construction that is examined. As expected, they are favoured in different national varieties, with poetic being the main collocate in the US and Canada, and lyrical in other inner-circle varieties (British, Irish, Australian and New Zealand English). Other results are less expected: in Indian and Sri Lankan English, the favoured collocate is eloquent rather than lyrical or poetic, and in Nigerian English, the main adjective in the data is stronger (as in [4]) a pattern that reflects the earlier, copular use of wax more directly than those found in other varieties.

(4) It is not strange because the Bible has warned us that evil men shall wax stronger and stronger and that iniquity shall abound. (GloWbE NG)

The results will be discussed against theories of World Englishes and previous research on phraseological variation across varieties of English.

References

Mukherjee, J. & Gries, S.Th. (2009). Collostructional nativisation in New Englishes: Verb-construction associations in the International Corpus of English. English World-Wide, 30(1), 27–51.

OED = Oxford English Dictionary Online. Oxford: Oxford University Press. Available at: www.oed.com.

Petré, P. (2012). General productivity: How become waxed and wax became a copula. Cognitive Linguistics, 23(1), 27–65.

L1 and L2 South African English past time reference

FULL PAPER

Bertus van Rooy & Ronel Wasserman (Universtity of Amsterdam, Netherlands)

Previous accounts of reference to situations in the past in Afrikaans English (ASAE) highlight the use of present tense forms, which is attributed to the tendency in Afrikaans not to mark past tense for past time forms (e.g. Watermeyer, 1996). However, no mention is made of the other transfer possibility, namely that the present perfect is used in contexts where other varieties of English would use the simple past. This is likely, since the Afrikaans past tense is morphosyntactically expressed only through the periphrastic form het+ge-V 'have+V_PastParticiple', having lost the inflected past tense of its Dutch ancestor. In view of the worldwide variability of the relationship of present perfect to simple past (Werner et al. 2016), this paper examines how the relationship is expressed in ASAE, in comparison to the transplanted variety of White South African English (WSAE). Parallel corpora of unedited written texts and their professionally edited versions (academic, instructional and popular writing, as well as news reportage) are analysed for these two varieties, to determine if unedited ASAE uses fewer morphosyntactically encoded past time references than WSAE in the unedited texts, and whether these texts are adjusted towards more past time reference by professional editors. Editorial adjustment helps to quantify the extent of present tense use for past time reference, and also indicates to what degree present perfect use is adjusted towards the simple past. The provisional results show a general trend of more editorial adjustments made to ASAE than to WSAE in this regard. Comparing the distribution of the non-perfect past and non-perfect present, ASAE uses the inflected past tense more frequently (25% of all non-perfects versus 16%. The editorial adjustments increase the simple past tenses even further in three of the four registers, though. Rather unexpectedly, the ASAE writers therefore use the present perfect proportionally less often than the simple past in comparison to the WSAE writers (10% to 14%), and editorial changes lead to a further increase in the simple past relative to the present perfect in three of the four registers as far as non-perfect forms are concerned. The results also show that ASAE uses the inflected past tense form proportionally more frequently than WSAE in the unedited forms already (25% versus 16% of all non-perfect aspects) in three of the four registers, leading to a further increase in inflected past tense forms and proportional decrease in present tense forms. These findings are very surprising from the perspective of cross-linguistic transfer, indicating that ASAE developed in the opposite direction than Afrikaans in comparison to

Dutch. Through qualitative analysis of concordance lines in the edited and unedited data, and comparison to available accounts for e.g. British English (BrE) (cf. Fuchs 2016, Werner 2014), we will endeavour to make sense of the unexpected development in ASAE, while also considering the possible degree of functional, if not quantitative, similarity between ASAE and WSAE.

References

Fuchs, R. (2016). The frequency of the present perfect in varieties of English around the world. In V. Werner, E. Seoane & C. Suárez-Gómez (Ed.), Re-assessing the Present Perfect (pp. 223-258). Berlin, Boston: De Gruyter Mouton.

Werner, V. (2014). The Present Perfect in World Englishes: Charting Unity and Diversity. PhD Thesis. Bamberg: University of Bamberg Press.

Werner, V., Seoane, E. & Suárez-Gómez, C. (2016). Re-assessing the Present Perfect. Berlin, Boston: De Gruyter Mouton.

Watermeyer, S. (1996). Afrikaans English. In V. de Klerk (Ed.), Focus on South Africa (pp. 99-148). Amsterdam: John Benjamins.

RHINE

MOSELLE

Thursday
28 May 2026

Thursday 28 May 2026

Poster Lightning Talks E011 • 10:00–10:30

Fifty Years of the Oxford Text Archive: digital research infrastructure then and now

POSTER

Martin Wynne (University of Oxford, United Kingdom)

The Oxford Text Archive celebrates its fiftieth anniversary in 2026. What can be learned from 50 years of curating corpora? Can we avoid making some of the same mistakes again? What will be the disruptive effects of new AI technologies have on repositories of language resources?

The Oxford Text Archive has been existence since 1976, with the history partially explored by Burnard (n.d.), Proud (1989), Pajares Tosca (2000), and Hockey (2004). continuously offering repository services for digital linguistic and literary resources. Over this period, at various times, the main focus of activity has changed in response to a variety of factors, including the needs of the community, changes in technology, local priorities, and funding opportunities. This focus has therefore alternated between archiving, collections development, research software development, user support, digitization, standards, resource creation, teaching and learning, local, national and international infrastructure. Highlights have included the Oxford Concordance Programme, an OCR service with the Kurzweil data entry machine (KDEM), the launch of the Text Encoding Initiative (Burnard 1988), and the construction of the British National Corpus by a consortium including OTA staff, and the curation of the Early English Books Online (EEBO) collection in TEI format.

In 1996 the OTA became one of the pillars of the Arts and Humanities Data Service (Burnard and Short, 1996), which ran until 2008, and at the same time as the downfall of the AHDS, the CLARIN pan-European research infrastructure started with similar aims (Váradi et al, 2008), with the OTA as one of the founding centres and coordinator of CLARIN in the UK. National infrastructure for the arts and humanities in the UK rebooted in 2021, with the OTA as a trusted repository as part of the ongoing Infrastructure for Digital Arts and Humanities (iDAH) programme. Collections development is firmly established as the key activity in the current period, and as the fiftieth anniversary looms, there are more than 70,000 items in the repository, with regular new deposits of corpora, lexical datasets and other digital resources.

A recent review of the successes and failures of the full period of the history of the OTA, including a horizon-scanning exercise to identify future opportunities and threats, has led

to certain conclusions. It is clear that language data is of enduring value for research; synchronic representative corpora might become outdated as snapshots of contemporary usage, and they become historical corpora. And while AI might be able to generate synthetic language data, this will not reduce the value and importance of authentic data. Current opportunities therefore include ongoing curation of historical datasets, and the strengthening of repository security to ensure the correct identification and curation of authentic language data. There is also scope for a focus on the creation of corpora, NLP tools and language models for lesser-resourced, historical, regional, specialized and minoritized languages and varieties. A current focus of activity is therefore the development of digital resources and AI skills training materials for the lesser-resourced languages and varieties in Britain and Ireland.

References

Burnard, Lou (1988), 'Report of Workshop on Text Encoding Guidelines', Literary and Linguistic Computing 3: 131–3.

Burnard, Lou, and Harold Short (1996), An Arts and Humanities Data Service, JISC [http://www.ahds.ac.uk/about/documents/ahds-feasibility-study.pdf]

Burnard, Lou (undated), 'Humanities Computing in Oxford: a Retrospective' [http://users.ox.ac.uk/~lou/wip/hcu-obit.txt]

Hockey, Susan (2004), 'The History of Humanities Computing', in A Companion to Digital Humanities, ed. Susan Schreibman, Ray Siemens, John Unsworth. Oxford: Blackwell, [http://www.digitalhumanities.org/companion /]

Proud, Judith K. (1989). The Oxford Text Archive. London: British Library Research and Development Report.

Pajares Tosca, Susana (2000), Report on the Humanities Computing Unit,[https://pendientemigracion.ucm.es/info/especulo/hipertul/HCUreport/HCUeng.htm]

Váradi, Tamas, Wittenberg, Peter, Krauwer, Steven, Wynne, Martin and Koskenniemi, Kimmo (2008). CLARIN: Common language resources and technology infrastructure. Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008.

Keyword analysis of linguistic features of British legal texts before and after Brexit: A corpus-based study

POSTER

Darya Katkova (Heidelberg University, Germany)

This study investigates how the linguistic structure of British legislative texts has changed over time and how EU law influenced British law before and after Brexit. The UK's shifting relationship with the European Union provides a unique opportunity to examine language change within a conservative and highly regulated register: legal discourse. Because legislative language typically changes very slowly, Brexit represents a rare moment when political and institutional transitions leave a visible trace in linguistic form. This poster presents

an overview of my PhD project, which investigates legal discourse, the characteristics of legal language, diachronic change in British legislative drafting across four dimensions: (1) lexical patterns, (2) syntactic and grammatical features, and (3) archaisms, (4) impersonal structures.

In line with these aims, the project addresses two research questions:

RQ1: How did the language of the EU law influence the British lexico-grammatical and syntactic features of legal documents during the EU membership?

RQ2: How did the linguistic structure of English legal documents change before, during and post-Brexit landscape?

The study draws on a diachronic corpus of UK Acts and Statutory Instruments from www.legislation.gov.uk, divided into three periods: pre-EU (1933–1993), EU membership (1993–2020), and post-Brexit (2020–2025). In my empirical research, I conducted a keyword and collocation analysis, and will subsequently conduct a concordance analysis of grammatical and syntactic features, archaisms, and impersonal structures. In my poster, I would like to highlight the analysis of keywords. This approach allows the study to capture how the language of UK legislation becomes more similar or different across institutional periods.

The analysis combines quantitative keyword extraction with qualitative linguistic interpretation. It examines:

1) positive and negative keywords,

2) types of legal terms based on online dictionaries.

3) keywords as markers of stylistic convention and carriers of substantive legal meaning.

Preliminary results show clear shifts in linguistic behaviour that correspond to political change. During EU membership, positive keywords strongly reflect EU-related terminology and integration-oriented concepts—for example, European, union, employees, tribunal, regulation, amended, involvement, and ballot. Negative keywords—such as children, art, persons, date, future, shares—tend to reflect everyday or domestically focused vocabulary. This pattern indicates the strong presence of EU discourse in UK legal drafting.

After Brexit, the trend reverses. Positive keywords highlight domestic governance, such as planning, licence, authority, immigration, sanctions, procurement, elections, voter, and local. Meanwhile, EU-related words such as European, union, employee, liability, and member appear among negative keywords, signalling a clear linguistic withdrawal from supranational framing.

Classifying positive keywords per period further shows a significant reduction in EU legal terms and workplace-related vocabulary, accompanied by a notable increase in institutional nouns (ministers, parliament, authority) and legislative verbs (granted, devolved, allocated, authorises), reflecting internal legal reform after Brexit.

Overall, the findings show that although legal discourse is traditionally stable, it undergoes systematic linguistic change when its institutional environment is restructured. The study contributes to research on language change in legal discourse, diachronic register variation, and the evolving dynamics of British legislative style before and after Brexit.

Online resources

1.EUR-Lex https://eur-lex.europa.eu/summary/glossary.html

2.GOV.UK: https://www.gov.uk/
3.Legislation.gov.uk: https://www.legislation.gov.uk/?Legislation.gov.uk
4.Plain English Campaign https://www.plainenglish.co.uk/a-to-z-of-legal-phrases.html#P
5.The Law Dictionary https://thelawdictionary.org/ .

References

Alasmary, A. A. (2019). Keywords in written academic legal texts: A corpus-derived list. International Journal of English Linguistics, 9(3), 40–50. Mellinkoff, D. (1963). The language of the law. Little, Brown & Co.

Scott, M. (2010). Problems in investigating keyness, or clearing the undergrowth and marking out trails. Corpora, 5(1), 85–104.

Stubbs, M. (2010). Three concepts of keywords. In Keyness in Texts (pp. 21–42).

Tiersma, P. M. (1999). Legal language. University of Chicago Press.

Power to the Corpus: Expected effect sizes and statistical power in multifactorial alternation studies

POSTER

Elen Le Foll (University of Cologne, Germany)

Statistical significance testing is widely used in corpus linguistics research (Buschfeld et al., 2024). As corpora have become ever larger, corpus linguists have traditionally focused on mitigating Type I errors (i.e. false positives), given that large sample sizes risk yielding statistically significant but practically insignificant results (see e.g. Brezina & Meyerhoff, 2014). However, multifactorial corpus analyses, particularly those relying on manual annotation, can also be prone to Type II errors (i.e. false negatives). These errors occur when small effects are reported as non-significant due to insufficient statistical power, often resulting from inadequate sample sizes in light of the number of predictors entered in any one model. Language being a complex phenomenon, it is typically modeled using numerous predictors and interactions, some of which may exhibit (very) small, yet real effects, that – to complicate matters further – may be subject to substantial variation. To tackle this thorny issue, this study attempts to estimate expected effect sizes for predictor variables commonly used in corpus-based multifactorial studies of morphosyntactic alternations. It is hoped that these expected effect sizes will enable us to conduct informed power analyses and calculate appropriate sample sizes for future corpus analyses of this kind.
The present study focuses on the calculation of expected effect sizes of continuous and categorical predictors in a highly-studied morphosyntactic alternation: the English dative alternation. To begin, a scoping review of multifactorial corpus studies was conducted. Only articles providing sufficient information to extract effect sizes for each coefficient estimate were retained. Meta-analytic methods are currently being applied to calculate average effect sizes, which will be reported alongside 95% confidence intervals. Based on the findings, it is recommended that effect sizes be routinely reported alongside statistical

significance to facilitate the interpretation and comparison of results. To enable future meta-analyses and promote reproducibility, annotated data and code should be made publicly available in trusted repositories, ideally following the FAIR principles (Wilkinson et al., 2016). By adopting these practices, the field of corpus linguistics can enhance the cumulative nature of its research and foster more informed and accurate conclusions. As the use of automated AI-driven data analyses in research increases, there is a real risk of automating large corpus-linguistic analyses and reporting statistically significant yet practically irrelevant findings, as observed in other disciplines (e.g. medicine, see Spick et al., 2025).

The results of the present study highlight the continued need to raise awareness about the significance (no pun intended) of effect sizes: despite numerous calls to do so (see e.g. Paquot & Plonsky, 2017; Wallis, 2020), these are not systematically reported and are frequently only very summarily interpreted. The average effect sizes computed in this study can be used to determine the minimum sample size required for the pre-registration of novel corpus analyses (Mak, 2024), a procedure that aims to improve the transparency, reproducibility, and ultimately trustworthiness of empirical research, but which has yet to be implemented in quantitative corpus-linguistic studies.

References

Brezina, V., & Meyerhoff, M. (2014). Significant or random?: A critical review of sociolinguistic generalisations based on large corpora. International Journal of Corpus Linguistics, 19(1), 1–28. https://doi.org/10.1075/ijcl.19.1.01bre

Buschfeld, S., Leuckert, S., Weihs, C., & Weilinghoff, A. (2024). How real is the quantitative turn? Investigating statistics as the new normal in linguistics. ICAME Journal, 48(1), 1–22. https://doi.org/10.2478/icame-2024-0001

Mak, M. H. C. (2024). Corpus linguistics will benefit from greater adoption of pre-registration: A novice-friendly split-corpus approach to pre-registration. Applied Corpus Linguistics, 4(3), 100111. https://doi.org/10.1016/j.acorp.2024.100111

Paquot, M., & Plonsky, L. (2017). Quantitative research methods and study quality in learner corpus research. International Journal of Learner Corpus Research, 3(1), 61–94. https://doi.org/10.1075/ijlcr.3.1.03paq

Spick, M., Onoja, A., Harrison, C., Stender, S., Byrne, J., & Geifman, N. (2025). Quantifying new threats to health and biomedical literature integrity from rapidly scaled publications and problematic research (p. 2025.07.07.25331008). medRxiv. https://doi.org/10.1101/2025.07.07.25331008

Wallis, S. (2020). Statistics in Corpus Linguistics Research: A New Approach. Routledge. https://doi.org/10.4324/9780429491696

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3(1), Article 1. https://doi.org/10.1038/sdata.2016.18

A Cross-linguistic Comparison of Speech Prosody and Rap Flows in American English and Cantonese

POSTER

Xingni Li (University of Oxford, United Kingdom)

Research on the relationship between language and music has examined how the linguistic prosody shapes musical forms within a speech community (e.g., Patel et al., 2006), providing empirical support for the long-standing intuition that the prosody of a culture's native language is reflected in its musical forms, and contributing to a broader understanding of the extent to which musical composition and perception draw on cognitive and neural mechanisms associated with language (Patel, 2008). However, little work has examined how speech prosody manifests in hybrid musical forms such as rap. Gilbers et al. (2020) investigated regional variation in speech prosody and rap flow across East Coast and West Coast African American English, but this line of inquiry has not yet been extended to a broader cross-linguistic perspective.

This study compares American English and Cantonese, two languages with contrasting rhythmic and word-prosodic typologies, to explore how prosodic characteristics are reflected in both speech and rap, and how linguistic prosody maps onto vocal musical expression. Specifically, the study addresses the following research questions:

(1) Do American English and Cantonese exhibit different rhythmic patterns in speech, and do these patterns shape their respective rap performance?
(2) Do American English and Cantonese exhibit different melodic patterns in speech, and do these patterns shape their respective rap performance?
(3) How are Cantonese tones transferred into rap?

The corpus consists of three interview clips of 10–15 seconds and two rap verses from each of four American English rappers and four Cantonese rappers. Rhythm is analysed using three vocalic interval-based metrics (%V, VarcoV, and nPVI), while melody is examined through two measures of pitch variability: pitch height variability (the coefficient of variation for deviation from mean pitch height) and pitch interval variability (the coefficient of variation for distances between consecutive pitch intervals). Cantonese tones are further examined through comparisons of tonal height, tonal slope value, and contour-tone slope direction between speech and rap.

The results show that, while American English and Cantonese exhibit clear typological differences in speech rhythm, these differences largely disappear in rap, primarily due to a reduction in nPVI in English and %V in Cantonese. No significant differences were found for either measure of pitch variability. In terms of tone, Cantonese tonal height relations and slope directions are largely preserved from speech to rap. At the same time, tonal pitch height undergoes a redistribution across registers, with non-low tones showing falling tendencies and contour tones becoming flatter.

These findings suggest that rap, as a global musical genre, imposes prosodic constraints

and aesthetic preferences on speech. At the same time, Cantonese rap appears to move toward prosodic conventions associated with American English rap. Overall, this study contributes to research on language and music by highlighting the role of musical practice in shaping the realisation of speech prosody in hybrid musical forms.

References

Gilbers, S., Hoeksema, N., De Bot, K., & Lowie, W. (2020). Regional variation in West and East Coast African-American English prosody and rap flows. Language and Speech, 63(4), 713-745.

Patel, A. D. (2008). Music, language, and the brain. Oxford University Press.

Patel, A. D., Iversen, J. R., & Rosenberg, J. C. (2006). Comparing the rhythm and melody of speech and music: The case of British English and French. The Journal of the Acoustical Society of America, 119(5), 3034–3047.

Politeness and Flirting E113 • 11:00–12:30

"I have to ask – do people just not flirt in Germany?": A cross-linguistic comparison of flirting strategies

FULL PAPER

Michelle Weckermann & Lena Scharrer (University of Augsburg, Germany)

The Netflix series Love is Blind seems to be a long runner in the dating show cosmos, having been expanded to no less than ten languages. Attracting many viewers worldwide, audiences appear to have noticed cultural differences in flirting behaviour, arguing for instance that Germans are not flirtatious, appearing clumsy and cold (Reddit, r/LoveIsBlindOnNetflix). This poses the question whether there are actual differences in flirting behaviour cross-linguistically, which this study sets out to explore. For this purpose, we compiled two corpora based on Love is Blind Germany and Love is Blind US, attempting to compare flirting behaviour in German and English. The German corpus consists of the first and thus far only season of the TV series (aired in 2025). For the English data set, we chose Season 8 in order to have data from a comparable time period.

Flirting, for the purposes of our study, is defined as a "strategic intentional action with the goal of signalling sexual and/or romantic interest in the interlocutor" (Scharrer & Weckermann, forthcoming). Moreover, flirting is inherently playful and involves a transgression, i.e., signalling more intimacy than is currently given (Scharrer & Weckermann, forthcoming; Speer, 2017). As there is thus far only limited linguistic research on flirting (see, for instance, Speer, 2017; Kiesling, 2013; Motschenbacher, 2020), it is no surprise that it has not yet been examined from a cross-linguistic perspective. German and English in general have, of course, garnered scholarly interest. Previous cross-linguistic research on German and English has shown, for instance, that Germans are more direct, self-oriented, and content-oriented in their communicative behaviour (see, for instance, House, 2006). This raises the question

whether this tendency is also reflected in a dating context, especially in dating shows. Previous studies have found that flirting employs a limited number of vehicles, including compliments, self-praise, imagined future, sexual innuendo, metalinguistic references, and humour (Scharrer & Weckermann, forthcoming; see also Speer, 2017; Kiesling, 2013; Mortensen, 2017). Both corpora were manually coded according to these vehicles and compared both qualitatively and quantitatively. We consider manual coding essential for this purpose, as these vehicles are not exclusively used in flirtatious contexts but also exist outside of flirtatious interactions. Self-praise, for instance, always involves a positive self-evaluation (either explicitly or implicitly; Dayter, 2018). However, it must signal a romantic/sexual interest, be playful, and involve the aforementioned transgression to qualify as flirting, for instance highlighting the speaker's own sexual prowess and/or desirability as a potential partner.

Preliminary results suggest that there is a difference in flirting behaviour between American and German participants. Generally speaking, Germans appear to flirt less frequently than Americans. Furthermore, there is an interesting difference in flirting vehicles used in both languages. Where Americans primarily rely on imagined future, Germans seem to signal their flirtatious intent mostly through sexual innuendo, rarely drawing on imagined future. Other vehicles like humour, self-praise, and compliments appear in both data sets.

References

Dayter, D. (2018). Self-praise online and offline: the hallmark speech act of social media? Internet Pragmatics, 1, pp. 184-203.

House, J. (2006). Communicative styles in English and German. European Journal of English Linguistics, 10(3), pp. 249-267.

Kiesling, S. F. (2013). Flirting and 'normative' sexualities. Journal of Language and Sexuality, 2(1), pp. 102-122.

Love is Blind (Germany) © Kinetic Content/Netflix (2025).

Love Is Blind (US) © Kinetic Content/Netflix (2025).

Mortensen, K. K. (2017). Flirting in online dating: Giving empirical grounds to flirtatious implicitness. Discourse Studies, 19(5), pp. 581-597.

Motschenbacher, H. (2020). Coming out – seducing – flirting: Shedding light on sexual speech acts. Journal of Pragmatics, 170, pp. 256-270.

Scharrer, L. & Weckermann, M. (2026). "You Wanna Flirt?" – "Let's Flirt": A Corpus-Based Analysis of Flirting. Journal of Pragmatics, 251, pp. 1-13.

Speer, S. A. (2017). Flirting: A decidedly ambiguous action? Research on Language and Social Interaction, 50(2), pp. 128-150.

How agent initiative shapes user politeness in cooking-related human–GenAI interactions

FULL PAPER

Christine Elsweiler (University of Innsbruck, Austria), Anna Ziegner (University of Innsbruck, Austria) & David Elsweiler (Unversity of Regensburg, Germany)

The question of whether users should interact politely with generative AI (GenAI) systems such as ChatGPT has sparked public debate, yet little is known about the politeness patterns users exhibit in such interactions. To date, user politeness towards GenAI systems has mainly been approached from a relatively narrow perspective, for instance, focussing on politeness markers such as please and thank you (Barko-Sherif et al. 2020; Yuan et al. 2024). Only few studies have examined user politeness in more depth (e.g., Tore 2025). In a prior study (AUTHOR(S) under review), we took a comprehensive approach to user politeness by analysing politeness profiles in a sub-set of Frummet et al.'s (2024) Cooking with Conversation data, featuring 30 task-oriented information-seeking conversations. The participants were interacting with a Wizard simulating an AI assistant, unbeknownst to them during the experiment, to complete a recipe consisting of several steps. Using an adapted version of Leech's (2014) classification scheme for politeness-sensitive speech acts, we derived four politeness clusters representing a continuum of politeness patterns: Hyperpolite, Polite and Engaged, Engagement-Seeking and Hyperefficient. At the polite end of the continuum, participants exhibit a high level of engagement and frequently use both face-threat mitigating and face-enhancing speech acts such as conventionally indirect requests as well as thanking, praising and acknowledgement of agent responses. Conversely, at the opposite end, participants display a task-oriented and efficient interactional style and only rarely employ face-enhancing strategies. This analysis was, however, limited to the passive condition, where the assistant only responded to explicit user requests.

In the present study, we extend the analysis to the active condition, where the assistant proactively provided background knowledge and assistance. We specifically aim to explore how the agent's active role affects the interactional style and politeness behaviour of participants compared to the baseline provided in the passive condition. We are currently annotating all politeness-relevant speech acts in the 25 conversations from the active condition, using the same classification scheme (Leech 2014) as for the conversations in the passive condition. A cluster analysis will be conducted to compare politeness profiles across the active and passive conditions. Preliminary findings show that, while generally the same politeness-sensitive speech acts are observed across conditions, their distribution seems to differ. The participants in the active condition exhibit higher engagement with the agent by asking more information-eliciting questions, such as Do I manually whisk or use a mixing machine? or What is brining?. Moreover, participants overall appear to employ more face-enhancing moves such as thanking the agent for the information provided and they use the politeness marker please more frequently. This suggests that the agent's proactive behaviour may engender more respectful and appreciative politeness behaviour on the part of the participants, which is characteristic of human-human interactions (Glas & Pelachaud

2015; Peters et al. 2005).

References

AUTHOR(S) (under review)

Barko-Sherif S., Elsweiler, D. & Harvey, M. (2020). Conversational agents for recipe recommendation. Proceedings of the 2020 Conference on Human Information Interaction and Retrieval, 73–82.

Frummet, A., Speggiorin, A., Elsweiler, D., Leuski, A., & Dalton, J. (2024). Cooking with conversation: Enhancing user engagement and learning with a knowledge-enhancing assistant. ACM Transactions on Information Systems, 42(5), 1–29.

Glas, N. & Pelachaud, C. (2015). Politeness versus perceived engagement: An experimental study. In Sharp, B. & Delmonte R. (Eds.), Natural language processing and cognitive science. Proceedings 2014 (no pagination). De Gruyter.

Leech, G. (2014). The pragmatics of politeness. Oxford: Oxford University Press.

Peters, C., Pelachaud, C., Bevacqua, E., Mancini, M., & Poggi, I. (2005). Engagement capabilities for ECAs. In Pelachaud, C., André, E., Kopp, S. & Ruttkay Z.M. (Eds.), Proceedings of workshop 13: Creating bonds with embodied conversational agents (no pagination). Utrecht University.

Tore, B. (2025). Speaking with AI: Gender-based variation in politeness and communication strategies. Masters thesis, Vytautas Magnus University. https://portalcris.vdu.lt/server/api/core/bitstreams/3635b4d4-fc9a-4181-8ee4-670c4998b515/content (accessed 17/09/2025).

Yuan, Y., Su, M., & Li, X. (2024). What makes people say thanks to AI. In Degen, H. & Ntoa, S. (Eds.), Artificial intelligence in HCI. 5th International Conference, AI-HCI 2024, held as part of the 26th HCI International Conference, HCII 2024, Washington, DC, USA, June 29 – July 4, 2024, proceedings, part I (pp. 131–149). Springer Nature Switzerland.

Politeness in Hong Kong business correspondence: Investigating genre development and nativization processes from a diachronic perspective

FULL PAPER

Carina Stick (Julius Maximilians University Würzburg, Germany)

Not only has the status of Hong Kong English (HKE) been questioned by many researchers, studies that have investigated HKE as a variety often also found features that were formerly assumed to be variety-specific to be genre-specific specific (Noël & Van der Auwera 2015). Consequently, diachronic data is needed to assess changes in one genre over time and in comparison with the superstrate to be able to distinguish between varietal and genre developments.

As this project is interested in more subtle, culturally induced changes rather than grammatical features (e.g. Edwards 2016), politeness is an interesting phenomenon. While most research on politeness in business correspondence has focused on requests (e.g. Del Lungo Camiciotti 2008), the fact that these often occur in combination with other moves, such as justifications (Kong 1998) or apologies (Blum-Kulka & Olshtain 1989), has been neglected to

date. Therefore, the project seeks to investigate the following questions:

- How has the use of requests and their accompanying moves in HK business correspondence changed from the 1900s to the 1960s? How does this tie in with the development of the genre as a whole?
- In the context of postcolonial Hong Kong: To what extent has the genre developed differently in the colony than in the colonizing country, based on an analysis of HK and British business correspondence from the 1960s to the 1990s? To what extent do the letters from Hong Kong display processes of nativization?
- Methodology: What approach and methods are most useful for the study of requests and their accompanying moves?

The study is based on a corpus of business letters sent from or to three Hong Kong based businesses (John Swire & Sons, Jardine Matheson & Co., the Hong Kong and Shanghai Banking Corporation (HSBC)). The corpus covers the time period from the 1900s to the 1990s, with a sub-corpus of 25.000 words for each decade. In addition, for the comparison of the genre's development in Hong Kong and the United Kingdom, a comparable corpus of British business correspondence was compiled, so that the overall size of the corpus amounts to 300.000 words.
As of now, all requests are identified by manually reading each letter in the corpus. However, the goal is to automate this process by creating an extensive list of words and phrases that can capture the majority of requests when conducting a corpus search. The requests are coded according to an adapted coding scheme which is based on Blum-Kulka and Olshtain's (1989) coding manual. So far, data from seven different time periods has been digitized, coded and analyzed.
Regarding the general development of the genre, in line with previous research (e.g. Del Lungo Camiciotti 2006) it is expected that a decrease in formality and an increase in directness in the formulation of requests will be observed. Concerning the development of the genre in HK in comparison with British business correspondence, diverging trends are expected from the 1960s on, as the nativization of HKE is assumed to have set in in the 1960s (Schneider 2007).

References

Blum-Kulka, S. and Olshtain, E. (1989). Cross-cultural pragmatics: Requests and apologies. Norwood, NJ: Ablex.

Del Lungo Camiciotti, G. (2008). Two polite speech acts from a diachronic perspective: Aspects of the realisation of requesting and undertaking commitments in the nineteenth century commercial community. In A. H. Jucker & I. Taavitsainen (Eds.), Speech acts in the history of English (pp. 115-131). Amsterdam: John Benjamins.

Del Lungo Camiciotti, G. (2006). From Your obedient humble servants to Yours faithfully: The negotiation of professional roles in the commercial correspondence of the second half of the

nineteenth century. In M. Dossena & I. Taavitsainen (Eds.), Diachronic perspectives on domain-specific English (pp. 153-172). Peter Lang.

Edwards, J. H. (2015). The deletion of /t, d/ in Hong Kong English. World Englishes, 35(1).

Kong, K. (1998). Are simple business request letters really simple? A comparison of Chinese and English business request letters. Text & Talk, 18(1), 103-141.

Noel, D. & van der Auwera, J. (2015). Recent quantitative changes in the use of modals and quasi-modals in the Hong Kong, British and American printed press. In P. Collins (Ed.), Grammatical change in English world-wide (pp. 437-464). Amsterdam: John Benjamins.

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

Benchmarking Large Language Models for Linguistic Research

FULL PAPER

Ondrej Tichy, Barbora Bulantova, Magdalena Titlbachova, Anna Marklova & Jiri Milicka (Charles University, Czech Republic)

Recent advances in Large Language Models (LLMs) have shown that they can increasingly replace traditional NLP techniques across a range of linguistic tasks, such as part-of-speech tagging (Lai et al., 2023) or orthographic normalization (Titlbachová, 2025), and can even approach or surpass human annotators in tasks such as speech-act classification (Koeva, 2024) or genre annotation (Kuzman, 2023). However, the results of previous studies are often difficult to replicate, given the wide range of factors influencing LLM output, its inherent stochasticity, and the breakneck lifecycle of the individual models. Consequently, state-of-the-art results, their systematic comparison and generalization remain challenging. In this paper, we propose a set of guidelines and Python scripts designed to make benchmarking LLMs on linguistic tasks more accessible, reproducible, and comparable. We also conduct several benchmarking experiments using this methodology to validate it and to identify best practices for applying LLMs to linguistic research, as well as to determine which current models perform best in specific tasks.

The proposed guidelines define input and output data structures for a variety of linguistic tasks, recommend parameter settings (e.g. temperature, top_p, chain-of-thought reasoning, retrieval-augmented generation), and outline how to interpret outputs such as self-reported confidence scores and token-level probabilities (Miller, 2024). The accompanying scripts enable researchers to (re)run tests on new tasks or new models and to generate comparable reports. While model fine-tuning is outside the scope of our framework, we support both zero- and few-shot prompting, allowing users to provide ground-truth data for evaluation and, optionally, as few-shot examples.

In our own tests, we will focus on tasks that have not been largely solved by NLP (avoiding e.g. PoS tagging in English) and that are commonly performed by empirical and more specifically corpus linguists. While most tasks target Present-Day English, we also investigate how LLMs handle low-resource languages and non-standardized varieties by including Czech and earlier stages of English. The selected tasks range from morphological and syntactic classification (e.g. identifying nominal number in Old English or the syntactic role of non-finite verbs in Present-Day English) to pragmatic annotation (e.g. contextual functions of like), semantic disambiguation, and historical spelling normalization.

We benchmark both major commercial models (e.g. ChatGPT, Gemini, Claude) and leading open-source or smaller models (e.g. gpt-oss, LLaMA, DeepSeek, Mistral), including different quantizations and configurations (leveraging the resources of the e-infra.cz research infrastructure). This enables evaluation not only of their performance but also of factors such as cost, accessibility, and data security, as well as testing claims such as smaller models outperforming larger ones on simple binary classifications (Kostina et al., 2025).

Our preliminary findings suggest that some linguistic tasks, such as text normalization and basic morphological classification, are already well-suited to LLM applications. In contrast, more complex tasks requiring extensive context and elaborate hierarchical categorization, such as discourse and pragmatic annotation (c.f. Ma et al., 2025), still fall significantly short of human performance.

References

Bollmann, M. (2019). A large-scale comparison of historical text normalization systems. In J. Burstein, C. Doran, & T. Solorio (Eds.), arXiv [cs.CL] (pp. 3885–3898). arXiv. https://doi.org/10.18653/v1/n19-1389

Bronsdon, C. (2025, March 29). A Complete Guide to LLM Benchmark Categories. Galileo.Ai. https://galileo.ai/blog/llm-benchmarks-categories

Dasgupta, I., Lampinen, A. K., Chan, S. C. Y., Sheahan, H. R., Creswell, A., Kumaran, D., McClelland, J. L., & Hill, F. (2022). Language models show human-like content effects on reasoning tasks. In arXiv [cs.CL]. arXiv. http://arxiv.org/abs/2207.07051

Duncan, D. (2024). Does ChatGPT have sociolinguistic competence? Journal of Computer-Assisted Linguistic Research, 8, 51–75. https://doi.org/10.4995/jclr.2024.21958

Hämäläinen, M., Säily, T., Rueter, J., Tiedemann, J., & Mäkelä, E. (2019). Revisiting NMT for normalization of early English letters. Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, 71–75. https://doi.org/10.18653/v1/w19-2509

Koeva, S. (2024). Large language models in linguistic research: The pilot and the copilot. Proceedings of the Sixth International Conference on Computational Linguistics in Bulgaria, 319–329. https://doi.org/10.47810/clib.24.35

Kostina, A., Dikaiakos, M. D., Stefanidis, D., & Pallis, G. (2025). Large Language Models for text classification: Case study and comprehensive review. In arXiv [cs.CL]. arXiv. http://arxiv.org/abs/2501.08457

Kuzman, T., Mozetič, I., & Ljubešić, N. (2023). ChatGPT: Beginning of an end of manual linguistic data annotation? Use case of automatic genre identification. In arXiv [cs.CL]. arXiv.

https://doi.org/10.48550/arXiv.2303.03953

Lai, V., Ngo, N., Pouran Ben Veyseh, A., Man, H., Dernoncourt, F., Bui, T., & Nguyen, T. (2023). ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In H. Bouamor, J. Pino, & K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023 (pp. 13171–13189). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.878

Ma, B., Li, Y., Zhou, W., Gong, Z., Liu, Y. J., Jasinskaja, K., Friedrich, A., Hirschberg, J., Kreuter, F., & Plank, B. (2025). Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges. In arXiv [cs.CL]. arXiv. http://arxiv.org/abs/2502.12378

Miller, R. (2024, September 16). Confidence Unlocked: A Method to Measure Certainty in LLM Outputs. Medium. https://medium.com/@vatvenger/confidence-unlocked-a-method-to-measure-certainty-in-llm-outputs-1d921a4ca43c

Opitz, J., Wein, S., & Schneider, N. (2025). Natural language processing RELIES on linguistics. Computational Linguistics (Association for Computational Linguistics), 51(3), 1–24. https://doi.org/10.1162/coli_a_00560

Sumanathilaka, D., Micallef, N., & Hough, J. (2025). GlossGPT: GPT for word sense disambiguation using few-shot chain-of-thought prompting. Procedia Computer Science, 257, 785–792. https://doi.org/10.1016/j.procs.2025.03.101

Titlbachová, M. (2025). Normalization of non-standardized varieties of Early English using AI language models [Charles University]. http://hdl.handle.net/20.500.11956/205439

Grammatical and Syntactic bias in Large Language Models FULL PAPER

Jiří Milička (Charles University, Czech Republic), Laura Alexis Janda (UiT The Arctic University of Norway) & Dominika Kováříková (Charles University, Czech Republic)

It is well known that artificial intelligence tends to exaggerate skewed distributions across various domains (O'Neil 2016, Manning 2022, Hall et al. 2022, Bubeck et al. 2023), but there has not yet been an investigation of the accuracy of grammatical distributions in the language produced by LLM-sourced chatbots. While there have been investigations into how well LLMs handle specific constructions and morphological forms (Weissweiler et al. 2023), no research to date has investigated grammatical bias in LLMs. Our research question is: Do LLMs reproduce the grammatical and syntactic distributions observed in human-generated data, or do they change those distributions?

We explore grammatical bias in LLMs using Czech, as it has rich morphology and there exists both the AI Koditex corpus (Milička et al. 2025) which is generated by LLMs, and a comparable corpus of human-written texts (Koditex, Zasina et al. 2018). We determine the distribution of grammatical categories using Universal Dependencies tagged texts (de Manerfe 2021). The bias in syntactic categories is explored in both Czech and English since we can utilize the AI Brown corpus (Milička et al. 2025, comparable to the BE 21 corpus, Baker 2023).

The distribution of grammatical categories is very skewed already in human-written texts.

For example, in Czech corpora, among the fourteen combinations of case and number obligatorily marked on noun phrases, the most frequent one (nominative singular) accounts for approximately 23.6% of all uses of nouns. Surprisingly, it turns out that almost all frontier models represented in AI Koditex (various versions of GPT, Claude, Gemini, DeepSeek, Llama) use the nominative singular at a lower frequency than human texts, typically around 21% or even below 20%. Other frequent combinations of case and number behave similarly. Conversely, the least frequent combinations (for example, dative plural) are somewhat more frequent in LLM-generated texts. The distribution of cases is therefore less skewed in LLM-generated texts than in original human-written texts and has higher entropy.

We explore how general this phenomenon is by examining also the verbal subsystem, and syntactic UD functors. While the scope of our investigation is limited to two languages (Czech and English), the robustness of our results is striking, since it suggests that LLMs effectively alter grammatical distinctiveness.

Given the influence that the language of LLMs has already and will have on human language in the future, this question is interesting not only from the perspective of machine learning but also from the perspective of linguistics (one might say, human linguistics). In this regard, it is a kind of forward diachrony: meaning not that we are examining what changes have occurred in language, but rather that we can estimate what changes will occur in language.

References

Baker, P. (2023). A year to remember? Introducing the BE21 corpus and exploring recent part of speech tag change in British English. International Journal of Corpus Linguistics, 28(3), 407–429.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv:2303.12712.

de Marneffe, M.-C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal dependencies. Computational Linguistics, 47(2), 255–308.

Hall, M., van der Maaten, L., Gustafson, L., Jones, M., & Adcock, A. (2022). A systematic study of bias amplification. arXiv:2201.11706.

Milička, J., Marklová, A., & Cvrček, V. (2025). AI Brown and AI Koditex: LLM-generated corpora comparable to traditional corpora of English and Czech texts. arXiv:2509.22996.

O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown.

Weissweiler, L., T. He, N. Otani, D. R. Mortensen, L. Levin, H. Schütze. 2023. Construction Grammar Provides Unique Insight into Neural Language Models. arXiv:2302.02178v1.

Zhang, Z., & Neill, D. B. (2016). Identifying significant predictive bias in classifiers. arXiv:1611.08292.

Zasina, A. J., Lukeš, D., Komrsková, Z., Poukarová, P., & Řehorková, A. (2018). Koditex: korpus diverzifikovaných textů. Praha. Retrieved 2025-09-26, from `https://wiki.korpus.cz/doku.php/cnk:koditex`.

The impact of L1-specific training data and writing topic on transformer-based CEFR classification

FULL PAPER

Christopher Cooper (Waseda University, Japan)

In recent years, high accuracy has been reported in studies that have classified English learner writing by CEFR level using BERT models (e.g., Schmalz & Brutti, 2022). This research has tended to use large learner corpora such as the EFCamDat (Geertzen et al., 2014) that are imbalanced in terms of the learners' L1 background. As the influence of the learner's L1 has been shown to influence second language learning in areas such as morpheme acquisition (Murakami & Alexopoulou, 2016), article usage (Crosthwaite, 2016), word order and verb tense (Shatz, 2016), its effect is worth investigating in text classification studies. In addition, previous research has not yet controlled for topic, with the same topics often being allowed to feature in both training and test sets. The current study aims to address these issues by answering the following research questions:

1. What is the difference in CEFR classification accuracy when a BERT model is finetuned on EFCamDat texts from learners of one nationality (Brazilian, Chinese, French, or Japanese) and tested on texts from the same and different nationalities within the same corpus?
2. How does classification accuracy across all nationality combinations change when training and test sets have no overlapping topics?
3. How does classification accuracy change when models finetuned on EFCamDat texts are tested on texts from a different learner corpus?

DistilRoBERTa models were finetuned using texts from the EFCamDat from four different L1 backgrounds (Brazil, China, France, Japan) and tested on unseen texts from the same backgrounds within the same corpus. Two experiments were conducted with the EFCamDat texts in a topic-controlled (no topics could feature in both the training and test sets) and a non-topic-controlled condition to test the effect of topic. The finetuned models were also tested on essays from the Write & Improve Corpus 2024 (Nicholls et al., 2024) with Japanese and Chinese L1 backgrounds. A mixed-effects logistic regression model was fit using the topic-controlled condition data with classification accuracy (correct/incorrect) as the dependent variable, training nationality and test nationality (and their interaction) as fixed effects, and topic and individual text as random effects.

The results indicated a clear effect of topic, with F-scores ranging from .921 to .959 in the non-topic-controlled condition and .382 to .542 in the topic-controlled condition. The random effects in the regression results also indicated substantial variation related to topic (SD = 4.03). There appears to be little effect of the learners' L1 background on classification accuracy, as most of the fixed effects in the logistic regression were not significant and models that were finetuned and tested on the same learner population did not show a clear pattern for higher F-scores. Finally, classification accuracy was lower when the models were tested on the Write & Improve data (F-scores = .306 to .422). Despite the low classification accuracy in both the topic-controlled condition and Write & Improve data, the models were able to distinguish between lower-level and higher-level texts in both cases, although exact

CEFR-level prediction accuracy was low.

References

Crosthwaite, P. (2016). A longitudinal multidimensional analysis of EAP writing: Determining EAP course effectiveness. Journal of English for Academic Purposes, 22, 166–178. https://doi.org/10.1016/j.jeap.2016.04.005

Geertzen, J., Alexopoulou, D., & Korhonen, A. (2014). Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCamDat). In R. T. Miller, K. Martin I., C. Eddington M., A. Henery, N. Marcos Miguel, A. Tseng M., A. Tuninetti, & D. Walter (Eds), Selected Proceedings of the 2012 Second Language Research Forum: Building Bridges between Disciplines (pp. 240–254). Cascadilla Press.

Murakami, A., & Alexopoulou, T. (2016). L1 influence on the acquisition order of English grammatical morphemes: A learner corpus study. Studies in Second Language Acquisition, 38(3), 365–401. https://doi.org/10.1017/S0272263115000352

Nicholls, D., Caines, A., & Buttery, P. (2024). The Write & Improve Corpus 2024: Error-annotated and CEFR-labelled essays by learners of English. Apollo - University of Cambridge Repository. https://doi.org/10.17863/CAM.112997

Schmalz, V. J., & Brutti, A. (2022). Automatic assessment of English CEFR levels using BERT embeddings. In E. Fersini, M. Passarotti, & V. Patti (Eds), Proceedings of the Eighth Italian Conference on Computational Linguistics CliC-it 2021 (pp. 293–299). Accademia University Press. https://doi.org/10.4000/books.aaccademia.10828

Shatz, I. (2016). Native language influence during second language acquisition: A large-scale learner corpus analysis. In M. Hirakawa, J. Matthews, K. Otaki, N. Snape, & M. Umeda (Eds), Proceedings of the Pacific Second Language Research Forum (PacSLRF 2016) (pp. 175–180).

Historical Linguistics [1] E313 • 11:00–12:30

An Annotated Corpus of Middle English Documents: New Solutions and New Questions

FULL PAPER

Merja Stenroos, Geir Bergstrøm, Takuya Takahashi & Kjetil V. Thengs (University of Stavanger, Norway)

This paper presents preliminary results of LiTra, a five-year project which will study low-frequency geographical variation in medieval English as a means to reconstruct earlier language and population history. The project focuses on late Middle English texts, which are characterized by extreme linguistic (including spelling) variation that to a large extent patterns geographically. This variation poses in itself a challenge for corpus searches. The first major questions addressed have therefore concerned the methodologies and tools for compiling and annotating a suitable corpus.

Pre-modern texts seldom connect to precise localities and are therefore often unsuitable

as evidence of geographical variation. The only substantial exception is documentary texts, including administrative writings such as accounts, affidavits, letters and wills. We propose to expand and develop an existing corpus, A Corpus of Middle English Local Documents (MELD), which now contains some 2,500 texts (ca 1 million words) localized throughout England. To improve the geographical resolution and the scope for studying low-frequency forms, it will be expanded to 3,500-4,000 texts.

As we wish to carry out searches at different levels of language, several levels of annotation are needed, including POS-tagging, semantic tagging and spelling unit analysis. Because of the variable spelling and morphology, the requirement for all these approaches is to first produce 'normalized' and lemmatized versions of the corpus, which can then be used to define searches and as a basis for further annotation. Until recently, the production of such versions would have been prohibitively time-consuming. Our first research question is therefore: how can we produce a fully searchable corpus of variation-rich historical texts within a reasonable timescale?

The texts contained in MELD are manually transcribed from manuscript images, a process that is in itself labour-intensive. The next step, normalization and lemmatization, would be even more so if carried out manually. Tools such as VARD (Baron & Rayson 2008) have improved efficiency considerably but still require a considerable level of manual input. Over the last fifteen years, numerous solutions, including tools based on neural networks and deep learning, have been developed to deal with variable historical languages (e.g. Kestemont et al. 2017); however, reaching a consistent high level of accuracy has been difficult (see e.g Hämäläinen et al. 2018). The advent of powerful LLM-based tools in recent years appears to have changed the situation dramatically; however, their use for corpus annotation has to be thoroughly tested in terms of consistency and scaleability, and reviewing the results requires care. On the other hand, it is expected that the latter process will in itself provide a research tool for a central task of the present project: identifying low-frequency patterns.

The paper addresses the questions to what extent LLM-based tools can be used to automatize 1) the transcription process, 2) the normalization and lemmatization of the corpus. The results are evaluated in terms of accuracy, consistency and overall time use, and preliminary linguistic findings are presented to illustrate the methodology.

References

Baron, A. & P. Rayson (2008). VARD2: a tool for dealing with spelling variation in historical corpora. In Postgraduate Conference in Corpus Linguistics. Birmingham: Aston University.

Hämäläinen, M., T. Säily, J. Rueter, J. Tiedemann & E. Mäkelä (2018). Normalizing early English letters to Present-Day English spelling. In B. Alex et al. (eds), Proceedings of the 2nd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, 87–96. Stroudsburg, PA: The Association for Computational Linguistics.

Kestemont, M., G. de Pauw, R. van Nie & W. Daelemans (2017). Lemmatisation for variation-rich languages using Deep Learning. Digital Scholarship in the Humanities 32(4): 797–815.

MELD = A Corpus of Middle English Local Documents. (2020-) Version 2017.1. University of

Stavanger. www.uis.no/meld

Latest publications by first author: Merja Stenroos (2025), The Geography of English in England. In Laura Wright and Raymond Hickey (eds), The New Cambridge History of English, Vol. I: Context, Contact and Development. Cambridge: Cambridge University Press. 99-129.

Conceptual Change during the Chemical Revolution: Air, Acid and Water in the Royal Society Corpus

FULL PAPER

Bach Phan Tat, Dirk Speelman & Dirk Geeraerts (KU Leuven, Belgium)

Cultural and scientific changes are often reflected in language (Degaetano-Ortlieb & Teich, 2022; Hoijer, 1948; Witherspoon, 1980), and semantic change literature often highlights how conceptual changes are encoded in lexical behaviour (Geeraerts, 1997). Focusing on English scientific prose, we trace the linguistic evidence of the conceptual reorganization of air, acid, and water in late-eighteenth to early-nineteenth-century chemistry. As high-frequency, topic-bearing lexemes, their co-occurrence profiles reflect the timeline over which oxygen displaced phlogiston and the corresponding changes in practice. Period-by-period patterns in the Royal Society Corpus reveal the chronology of the phlogiston-oxygen paradigm shift and the linguistic imprint it left in English scientific writing.

The research question is: During the phlogiston-oxygen paradigm shift, how did core terms like air, water and acid shift in their conceptual elaboration?

We use SynFlow (Phan-Tất, 2025), a method based on syntactic co-occurrence and Jensen-Shannon Divergence (JSD) to model the diachronic distributional shifts of two syntactic slot-fillers of air, acid and water across successive 5-year periods (technical details and package descriptions are covered in the accompanying software demonstration). The data is drawn from a subset (1750 – 1819, filtered by the topics of Chemistry 1, 2, Biochemistry (Menzel et al., 2021)) of the Royal Society Corpus 6.0 Open (Fischer et al., 2020). Inspired by Davies (2025), we adopted an AI-assisted interpretation workflow in which ChatGPT generated candidate qualitative interpretations of the quantitative output which were then independently verified with established narrative literature from the history and philosophy of science (e.g., Blumenthal & Ladyman, 2017; Boantza, 2013; Cavendish, 1766; Chang, 2011; Conant, 1950; Dalton, 1805; Henry, 1803; Kuhn & Hacking, 2012; McEvoy, 2015; Stewart, 2012). Although ChatGPT sometimes produced unsupported claims, it was often useful, especially with its recent 'Thinking' mode.

For air, in the 1750s-60s, high JSD reflects an exploratory period with the rise of a new set of descriptors (fixed, inflammable), consistent with phlogiston theory's consolidation-before-crisis stage. JSD drops in the 1770s–80s as the vocabulary usage temporarily stabilized during a period of debate between the main rival theories. JSD rises again in the 1790s as phlogiston terms recede yet persist as an epistemic object into the 1810s. Verbs in the oblique construction (e.g., expose to the air) move from passive exposure (tarnish, expose) to reagent-style operations (combine, unite, convert) to measurement and control (weigh,

confine, exhaust). Water shows a parallel shift from gas washing and eudiometric practice by volume to gravimetry and precise composition, then to roles as reactant and oxidant. Acid moves from source/salt/strength labels to oxygen-based composition and the -ic/-ous scheme. Together these trajectories show how distributional profiles trace both chronology and practice as chemistry reoriented from phlogiston theory to oxygen theory.

The contribution of this study is two-fold:

First, we analyse three central chemical concepts during the phlogiston-oxygen paradigm shift and show how the shift is reflected in scientific language using SynFlow, a linguistically-grounded, bottom-up method with a user-friendly package.

Second, we evaluate an AI-assisted interpretation workflow, noting its benefits and limitations (i.e., hallucination) for non-specialists in the history of science.

References

Blumenthal, G., & Ladyman, J. (2017). The development of problems within the phlogiston theories, 1766–1791. Foundations of Chemistry, 19(3), 241–280. https://doi.org/10.1007/s10698-017-9289-0

Boantza, V. D. (2013). The Rise and Fall of Nitrous Air Eudiometry: Enlightenment Ideals, Embodied Skills, and the Conflicts of Experimental Philosophy. History of Science, 51(4), 377–412. https://doi.org/10.1177/007327531305100401

Cavendish, H. (1766). XIX. Three papers, containing experiments on factitious air. Philosophical Transactions of the Royal Society of London, 56, 141–184. https://doi.org/10.1098/rstl.1766.0019

Chang, H. (2011). The Persistence of Epistemic Objects Through Scientific Change. Erkenntnis, 75(3), 413–429. https://doi.org/10.1007/s10670-011-9340-9

Conant, J. B. (1950). The Overthrow of the Phlogiston Theory: The Chemical Revolution of 1775-1789.

Dalton, J. (1805). On the Absorption of Gases by Water and other Liquids.

Davies, M. (2025). AI/LLM integration with the corpora from English-Corpora.org. https://www.english-corpora.org/ai-llms/

Degaetano-Ortlieb, S., & Teich, E. (2022). Toward an optimal code for communication: The case of scientific English. Corpus Linguistics and Linguistic Theory, 18(1), 175–207. https://doi.org/10.1515/cllt-2018-0088

Fischer, S., Knappen, J., Menzel, K., & Teich, E. (2020). The Royal Society Corpus 6.0: Providing 300+ Years of Scientific Writing for Humanistic Study.

Geeraerts, D. (1997). Diachronic Prototype Semantics: A Contribution Historical Lexicology (D. Geeraerts, Ed.). Oxford University Press. https://doi.org/10.1093/oso/9780198236528.003.0001

Henry, W. (1803). III. Experiments on the quantity of gases absorbed by water, at different temperatures, and under different pressures. Philosophical Transactions of the Royal Society of London, 93, 29–274. https://doi.org/10.1098/rstl.1803.0004

Hoijer, H. (1948). Linguistic and Cultural Change. Language.

Kuhn, T. S., & Hacking, I. (2012). The structure of scientific revolutions (Fourth edition). The

University of Chicago Press.

McEvoy, J. G. (2015). Gases, God and the balance of nature: A commentary on Priestley (1772) 'Observations on different kinds of air'. Phan-Tất, B. (2025). SynFlow [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.17414457

Stewart, J. (2012). The Reality of Phlogiston in Great Britain. HYLE–International Journal for Philosophy of Chemistry, 18.

Witherspoon, G. (1980). Language in Culture and Culture in Language. International Journal of American Linguistics, 46(1), 1–13. https://doi.org/10.1086/465623

Expanding the construct of grammatical complexity: a case for grammatical diversity in writing

FULL PAPER

Christian Holmberg Sjöling (Luleå University of Technology, Sweden) & Taehyeong Kim (Northern Arizona University)

Grammatical/syntactic complexity has received much attention with regards to language assessment (e.g., Lu, 2017), proficiency (e.g., Biber et al., 2016), and development (e.g., Biber et al., 2011) within the framework of Corpus Linguistics (CL). It has also been researched extensively in relation to, for example, register (e.g., Qin & Zhang, 2023) and production modality (e.g., Lintunen & Mäkilä, 2014). In most such studies, it is common for researchers to apply different automatic tools like the L2 Syntactic Complexity Analyzer (L2SCA) (Lu, 2010) or the Biber Tagger (Biber, 1988) to name a few.

Most of these programmes, and the studies conducted with them, employ frequency-based measures of grammatical/syntactic complexity, i.e., the occurrence of a particular feature is counted based on how often it occurs in an individual text belonging to, for example, a specific register and/or student population. Then, these frequency counts are used to draw conclusions about which features are characteristic of higher grades in high-stakes language tests (e.g., Kyle & Crossley, 2018) or differences between spoken and written registers (Biber et al., 2021). However, few programmes for automatic analysis of texts, if any, contain a linguistically interpretable measure of what we call grammatical diversity (see also Jarvis, 2013).

In this paper, we build on the frequency-based approach to tap into a complementary dimension of grammatical complexity that we term grammatical diversity. To explore the proposed construct, we have developed a measure in Python that builds on tagging initially carried out with the Lexicogrammatical Tagger (Kyle et al., 2025).

The measure counts the number of unique grammatical complexity features for every moving five-sentence window of a text (i.e., for sentence 1–5, 2–6, 3–7 and so on) and produces a final value of the moving average of unique grammatical complexity features

for all five-sentence windows of a text by dividing the total number of unique features per five-sentence window with the number of five-sentence windows. The final value is larger if a writer uses a variety of grammatical complexity features within a five-sentence window (i.e., exhibiting greater grammatical diversity), while it is lower if a writer repeats or use fewer grammatical complexity features within a five-sentence window (i.e., showing less grammatical diversity). The final value is calculated separately for clausal features and phrasal features to maintain linguistic interpretability (see Egbert et al., 2020; Biber et al., 2025). This approach also accounts for the finite number of grammatical features that writers eventually must repeat.

The measure is applied in a pilot study where a sample of 7,702 L2 texts from the EF-Cambridge Open Language Database (EFCAMDAT) (Geertzen et al., 2014; Huang et al., 2014) is analysed. Firstly, diagnostics were run to establish window size, then, a minimally sufficient approach from Staples et al. (2023) was used to determine if there was a difference of grammatical diversity between different proficiency levels in the corpus. The findings show a steady developmental increase across all proficiency levels for both phrasal and clausal diversity. The computation of the measurement and the findings are discussed further.

References

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

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

Biber, D., Gray, B., & Staples, S. (2016). Predicting patterns of grammatical complexity across language exam task types and proficiency levels. Applied Linguistics, 37(5), 639–668.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (2021). Grammar of spoken and written English. John Benjamins. [Previously published in 1999 by Longman].

Geertzen, J., Alexopoulou, T., & Korhonen, A. (2014). Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCamDat). In R.T. Millar, K.I. Martin, C.M. Eddington, A. Henery, N.M. Miguel, & A. Tseng (Eds.), Selected proceedings of the 2012 Second Language Research Forum (pp. 240–254). Somerville, MA: Cascadilla Proceedings Project.

Huang, Y., Geertzen, J., Baker, R., Korhonen, A., & Alexopoulou, T. (2017). The EF Cambridge Open Language Database (EFCAMDAT): Information for users (pp. 1–18).

Jarvis, S. (2013). Capturing the diversity in lexical diversity. Language learning, 63, 87–106.

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.

Kyle, K., Biber, D., Sung, H., Reppen., R., & Egbert, J. (2025). Lexicogrammatical Tagger. [computer software]. https://github.com/kristopherkyle/LxGrTgr .

Lintunen, P., & Mäkilä, M. (2014). Measuring syntactic complexity in spoken and written learner language: Comparing the incomparable? Research in Language, 12(4), 377–399.

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

Lu, X. (2017). Automated measurement of syntactic complexity in corpus-based L2 writing research and implications for writing assessment. Language Testing, 34(4), 493–511.

Staples, S., Gray, B., Biber, D., & Egbert, J. (2023). Writing Trajectories of Grammatical Complexity at the University: Comparing L1 and L2 English Writers in BAWE. Applied Linguistics, 44(1), 46–71.

Qin, W., & Zhang, X. (2023). Do EFL learners use different grammatical complexity features in writing across registers? Reading and Writing, 36(8), 1939–1967.

The Effect of Complexity on the Choice of a Complement Structure in a Subject Position. A Corpus-Based Study

FULL PAPER

Iverina Ivanova (Goethe University, Frankfurt am Main, Germany)

In this study, I investigated the Complexity Principle (CP), which posits that there is a correlation between syntactic explicitness and cognitive complexity, with the aim to determine whether complexity plays a role for the choice between finite (that-cl) and non-finite (to-inf) complement clauses in a subject position (Rohdenburg 1996, p.151). To avoid the collostruction effect (i.e., strong verb-structure associations), which according to Gries & Stefanowitsch (2003) and Gries (2005) can override other factors, such as syntactic priming, I focused on the distribution of the complement structures outside the VP predicate. To test the CP, I searched the English Gigaword corpus (Graff & Cieri 2003) and extracted 17,743 example sentences containing that and to-inf complement clauses as subjects followed by a VP predicate headed by the verb "be". Each complement clause was automatically annotated for the presence or absence of a predefined complexity predictor using spaCy and its transformer-based model en_core_web_trf. The aim was to determine which syntactic complexity predictors increase the probability of the syntactically more transparent clause (that-cl) and to estimate their individual strength effect, as well as their interaction effect by using Bayesian regression models (Bürkner 2020). The analyzed complexity predictors include: negation, passivization, modal verbs, noun modification, adverbial adjuncts, supplements, and the presence of embedded complement or adverbial clauses. For all these predictors, there is empirical evidence showing that their presence makes the clause not only longer, but also more difficult to process due to the increased memory demand and interpretation effort (e.g., Dudschig & Kaup 2020; Mack et al. 2013; Tsiamtsiouris & Cairns 2013).

In addition, Roland et al. (2012) found that processing is facilitated if the words occurring in a particular context are semantically similar to each other. To examine whether semantic cohesion influences the selection of a complement structure, I measured the degree of semantic relatedness among the content words in each complement clause using BERT embeddings. This allowed me to assess whether the words in the clause cluster together or are scattered in the embedding space. Finally, the average cohesion scores for the two types of complement clauses (that-clauses and to-inf. clauses) were computed and compared. The results from this study show that the presence of syntactic complexity factors increases

the probability of the finite complement clause. These results not only provide corpus-based evidence to Rohdenburg's claim (1995) that the internal structural complexity is an inherent characteristic of finite complement structures, but they can also be used to fine-tune a small generative language model (SLM) on the strongest predictors of that-clauses, with the aim to improve the functional knowledge of a generative LM by augmenting it with the complexity factors which proved to have the strongest effects on the choice of a complement structure.

References

Bürkner, P.-Ch. (2020). brms. Bayesian Regression Models using 'Stan'. https://cran.r-project.org/web/packages/brms/index.html

Dudschig, C., & Kaup, B. (2020). Can we prepare to negate? Negation as a reversaloperator. Journal of Cognition, 3(1), Article 32. https://doi.org/10.5334/joc.119

Graff, D., & Cieri, Ch. (2003). English Gigaword LDC2003T05. Web Download. Philadelphia: Linguistic Data Consortium

Gries, S. Th., & Stefanowitsch, A. (2003). Collostructions: Investigating theinteraction of words and constructions. International Journal of Corpus Linguistics, 8(2), 209 - 243.

Gries, S.Th. (2005). Syntactic Priming: A Corpus-based Approach. Journal of Psycholinguistic Research, 34 (4). https://doi.org/10.1007/s10936-005-6139-3

Mack, J.E., Meltzer-Asscher, A., Barbieri, E., & Thompson, C.K. (2013). Neural correlates of processing passive sentences. Brain Sci., 3(3),1198–1214. https://pmc.ncbi.nlm.nih.gov/articles/PMC4061884/

Rohdenburg, G. (1995).On the replacement of finite complement clauses by infinitives in English. English Studies, 76(4), 367–388.

Rohdenburg, G. (1996). Cognitive complexity and increased grammatical explicitness in English. Cognitive Linguistics, 7 (2), 149–182.

Roland, D., Yun, H., Koenig, J.-P., & Mauner, G. (2012). Semantic similarity, predictability, and models of sentence processing. Cognition, 122 (3), 267-279. https://doi.org/10.1016/j.cognition.2011.11.011

Tsiamtsiouris, J., & Cairns, H.S. (2013). Effects of sentence-structure complexity on speech initiation time and disfluency. Journal of Fluency Disorders, 38(1), 30-44. https://doi.org/10.1016/j.jfludis.2012.12.002

Does self-initiated reading and gaming affect intermediate-level L2 student writers' grammatical complexity?

FULL PAPER

Tove Larsson (Northern Arizona University, United States of America), Henrik Kaatari (University of Gävle, Sweden), Ying Wang (Karlstad University, Sweden), Pia Sundqvist (University of Oslo, Norway), Taehyeong Kim (Northern Arizona University, United States of America) & Douglas Biber (Northern Arizona University, United States of America)

In countries like Sweden, middle-school and high-school students are increasingly exposed

to English outside the classroom through self-initiated, Extramural English (EE) activities (e.g., gaming, social media). In line with usage-based theories (Ellis et al., 2016), we would expect that the type of EE engagement has an effect on students' output. While this hypothesis has not been systematically tested in the context of EE, there is some supporting evidence in the literature, in particular for EE gaming and EE reading: EE gaming (which tends to expose students to lexically varied, oral/spoken input) has been shown to positively affect students' vocabulary size (e.g., Sundqvist & Wikström, 2015), while EE reading (providing written/literate input) is found to positively affect students' noun phrase complexity (Kaatari et al., 2023). However, most previous research has examined EE activities in isolation, overlooking potential interaction effects from engaging in multiple activities simultaneously. The present study zooms in on grammatical complexity to test the hypothesis that the type of EE input students receive will affect their output, while taking learners' broader EE profiles into account. Grammatical complexity is defined as the addition of optional structural elements to simple phrases and clauses (Biber et al., 2022). It offers a helpful framework for analyzing possible effects of EE on students' output in that different groups of complexity features are associated with oral vs. literate registers: 'Oral' features (e.g., finite adverbial clauses, verb + that-complement clauses) are more frequent in oral/spoken registers, and 'literate' features (e.g., attributive adjectives, prepositional phrases) are associated with literate/written registers (Biber & Larsson, 2025).

Using the Swedish Learner English Corpus (Kaatari et al., 2024), we focus specifically on the possible effects of EE gaming and EE reading. We would expect increasing literate/written input through EE reading to lead to more literate/written-like complexity features, following Kaatari et al. (2023). For EE gaming, our analysis will be exploratory, but given the spoken elements of many online games (e.g., Reinhardt, 2019), we expect that more gaming will lead to more oral/spoken-like complexity features. We ask the following research questions:

1. To what extent does the time students spend on EE reading and EE gaming help predict frequency of phrasal vs. clausal complexity features in their writing?
2. Are there differences across profiles (e.g., high reading and high gaming) in students' use of phrasal vs. clausal complexity features?

Linear regression results show that both EE reading and EE gaming have a statistically significant positive effect on phrasal complexity, albeit with limited variance explained by the model (R2: 0.03). No effect of either EE activity was noted for the clausal features. Regarding RQ2, we looked at students whose EE reading/gaming time was in the top/bottom 30th percentile. We saw that the high gaming-high reading group had consistently higher frequencies for phrasal complexity with a low-to-medium effect size (d: .58, .21, .15). No main effects of the profiles were noted for clausal complexity. Partially contrary to our hypothesis, we thus conclude that both EE activities have a positive effect on phrasal complexity.

References

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

Biber, D., & Larsson, T. (2025). Accounting for the entire system of complexity features: Evidence for general oral versus literate grammatical complexity dimensions. Corpus Linguistics and Linguistic Theory. Available in early online access: https://doi.org/10.1515/cllt-2025-0017

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. Wiley-Blackwell.

Kaatari, H., Larsson, T., Wang, Y., Acikara-Eickhoff, S., & Sundqvist, P. (2023). Exploring the effects of target-language extramural activities on students' written production. Journal of Second Language Writing, 62, 101062.

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.

Reinhardt, J. (2019). Gameful second and foreign language teaching and learning: Theory, research, and practice. Palgrave Macmillan/Springer International Publishing.

Sundqvist, P., & Wikström, P. (2015). Out-of-school digital gameplay and in-school L2 English vocabulary outcomes. System, 51, 65–76.

ASR and Multimodal AI [1] E113 • 14:00–16:00

Fine-Tuning ASR for Corpus Linguistics: Singapore English FULL PAPER

Steven Coats (University of Oulu, Finland), Carmelo Alessandro Basile (Sorbonne Nouvelle University, France), Cameron Morin (Paris-Cité University, France) & Robert Fuchs (University of Bonn, Germany)

This paper presents recent advances in adapting large Automatic Speech Recognition (ASR) models for corpus-linguistic research on Singapore English, a rapidly endonormativizing variety (Leimgruber, 2013; Schneider, 2007; Tan, 2014) characterized by distinctive phonetic and grammatical features. Building on recent work in developing domain-specific ASR systems, we investigate how fine-tuned versions of Whisper (Radford et al., 2022) can improve transcription accuracy for Singapore English podcasts and conversational speech. Singapore English differs markedly from British and American English in pronunciation, syntax, and discourse-pragmatic use of particles such as lah, lor, sia, and kena. Although general-purpose ASR systems like Whisper perform well on global English varieties, they struggle with these features. The availability of the National Speech Corpus (NSC; Koh et al., 2019) has made it possible to fine-tune ASR models on locally representative data. Previous work has demonstrated improvements through large-scale fine-tuning, notably the MERaLiON model (Wang et al., 2025). However, the potential benefits of smaller, targeted fine-tuning for linguistic research remain underexplored.

We fine-tuned OpenAI's Whisper models on 895 hours of conversational Singapore English, comprising 871 distinct speakers, extracted from Part 3 of the NSC (1.26 million speaker

turns, 9.3 million words). We created six models trained on 30-second segments of speech, one for each of the model sizes tiny, base, small, medium, large-v2, and large-v3. In addition, we trained a large-v3 model on shorter segments comprising individual speaker turns. We compared our models with the six default Whisper variants, several out-of-the-box OWSM models from ESPnet (Peng et al., 2025), and the 10-billion-parameter MERaLiON-2-ASR (Wang et al., 2025), finding that our best-performing fine-tuned model exhibits similar perfomance on in-domain data to MERaLiON-2, with a much smaller computational overhead. Our results suggest that near-domain fine-tuning on moderately sized datasets can yield substantial gains for specific linguistic applications without requiring large compute budgets.

To assess the models' linguistic utility, we examined their transcription of distinctive SgE discourse particles. MERaLiON-2 and our fine-tuned models consistently outperformed baseline Whisper systems and OWSM models in recognizing lah, lor, sia, and kena, as well as in retaining repetitions, false starts, and self-repairs, features that are of interest in corpus-based analyses of interactional phenomena and pragmatic meaning in Singapore English. Overall, our experiments demonstrate that fine-tuning Whisper on Part 3 of the NSC substantially improves ASR performance for Singapore English. For corpus linguistics, our model will facilitate the automatic creation of a more accurate and linguistically rich version of the YouTube Corpus of Singapore English Podcasts (YCSEP; Coats et al., 2025). By combining scalable ASR fine-tuning with sociolinguistic insights, this study contributes to both applied speech technology and the systematic study of World Englishes.

References

Coats, Steven and Dana Roemling. (2025). The Corpus of Recorded Investigative, Media, and Evidence-based Proceedings. In Annamária Fábián and Igor Trost (eds.), Impulses and Approaches to Computer-Mediated Communication: Proceedings of the 12th International Conference on Computer Mediated Communication and Social Media Corpora for the Humanities, 45-49. Bayreuth, Germany: University of Bayreuth.

Coats, Steven, Carmelo Alessandro Basile, Cameron Morin and Robert Fuchs. (2025). The YouTube Corpus of Singapore English Podcasts. English World-Wide.

Coats, Steven. (2025). An automatic pipeline for processing streamed content: New horizons for corpus linguistics and phonetics. In Louis Cotgrove, Laura Herzberg, and Harald Lüngen (eds.), Exploring digitally-mediated communication with corpora: Methods, analyses, and corpus construction, 257–274. Berlin: De Gruyter Brill.

Morin, Cameron and Steven Coats. (2025). Double modals in Australian and New Zealand English. World Englishes 44(3), 415–438.

Coats, Steven. (2025). 'What the X' in Anglophone government meetings: Areal distribution, emotionality, and euphemism. Lingua 321.

Coats, Steven, Chloé Diskin-Holdaway, and Debbie Loakes. (2025). Regional distribution of the /el/-/æl/ merger in Australian English. In Yves Scherrer, Tommi Jauhiainen, Nikola Ljubešić, Preslav Nakov, Jorg Tiedemann, and Marcos Zampieri (eds.), Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects.

Towards a taxonomy of AI transcription and human correction: Pragmatic and speech production phenomena in the Corpus of Spoken Maldivian English (CoSpoMalE)

FULL PAPER

Julia Degenhardt (Justus Liebig University Giessen, Germany & Augsburg University, Germany) & Eliane Lorenz (Justus Liebig University Giessen, Germany)

Recent advances in Generative Artificial Intelligence (GenAI) have made speech-to-text models increasingly accessible for corpus-based linguistic research (Weilinghoff, 2025). While these tools substantially reduce transcription time (Kuhn et al., 2024), they often reshape spontaneous speech into standardized written form, omitting or smoothing phenomena that are vital for pragmatic and discourse analysis, such as pauses, false starts, repetitions, repairs, and phatic expressions (O'Connor Russell et al., 2024). This loss of interactional detail poses challenges for studies on spoken language, and risks obscuring linguistic cues that provide valuable information, especially for pragmatic and discourse-analytic research.
The present study uses the pilot version of the Corpus of Spoken Maldivian English (CoSpoMalE) – the first documentation of spoken English usage of speakers in their 20s and 30s residing in the Maldives, and thus of a population whose English language use is heavily influenced by code-switches between English and Dhivehi to address the following research questions:
1) How accurately can GenAI-based transcription tools capture corpus-pragmatic and fluency-related phenomena in code-switching spoken data?
2) How can human post-edits inform a taxonomy for improved AI-human transcription workflows?
To answer these research questions, we used noScribe (Dröge, 2024) to transcribe the CoSpoMalE recordings, which were subsequently reviewed and – if necessary – corrected by speakers of Maldivian English. The two versions of the transcripts were then compared to identify possibly systematic differences between AI-generated and human-corrected output.
This comparison shows that the GenAI-based transcriptions tend to omit false starts, merge repetitions, and delete or correct code-switches, which goes in hand with broader findings on the limitations of automatic speech recognition (ASR) in processing spontaneous and multilingual speech (Mustafa, 2022; Qiao et al., 2021; Weilinghoff, 2025). To address these discrepancies, this study has two aims, namely to provide an error index of how AI-transcription and human-correction differ in our data set and to propose a taxonomy of correction types that differentiates (1) surface corrections (orthographic, segmentation), (2) interpretive corrections (contextual inference, code-switching) and (3) pragmatic restorations (pauses, overlaps, fluency and discourse markers). Building on existing approaches to spontaneous-speech annotation (Candido Junior et al., 2023), this taxonomy offers a framework for hybrid transcription workflows that preserve pragmatic richness while still profiting from GenAI efficiency.
Hence, this study illustrates how human expertise and GenAI transcriptions complement

each other in current corpus linguistics research. It offers a linguistically grounded framework for integrating automation without compromising analytic depth, which is particularly relevant when working with varieties of English in which code-switching practices are an integral part of everyday communication.

References

Candido Junior, A., Casanova, E., Soares, A., de Oliveira, F. S., Oliveira, L., Corso Fernandes Junior, R., Peixoto Pinto da Silva, D., Gorgulho Fayet, F., Baldissera Carlotto, B., Stefanel Gris, L. R., & Aluísio, S. M. (2023). CORAA ASR: A large corpus of spontaneous and prepared speech manually validated for speech recognition in Brazilian Portuguese. Language Resources and Evaluation, 57, 1139–1171. https://doi.org/10.1007/s10579-022-09621-4 .

Dröge, K. (2024). noScribe. AI-powered audio transcription (Version 0.6.2) [Computer software].

Kuhn, K., Kersken, V., Reuter, B., Egger, N., & Zimmermann, G. (2024). Measuring the accuracy of automatic speech recognition: Word error rate and beyond. ACM Transactions on Accessible Computing, 16(4), Article No. 25. https://doi.org/10.1145/3636513 .

Mustafa, M. B. (2022). Code-switching in automatic speech recognition. Applied Sciences, 12, Article No. 9541. https://doi.org/10.3390/app121995410 .

O'Connor Russell, S., Gessinger, I., Krason, A., Vigliocco, G., & Harte, N. (2024). What automatic speech recognition can and cannot do for conversational speech transcription. Research Methods in Applied Linguistics, 3(3), Article No. 100163. https://doi.org/10.1016/j.rmal.2024.100163 .

Qiao, Y., Zhou, W., Kerz, E., Schlüter, R. (2021). The Impact of ASR on the Automatic Analysis of Linguistic Complexity and Sophistication in Spontaneous L2 Speech. Proceedings of Interspeech 2021, 4453–4457. https://doi.org/10.21437/Interspeech.2021-1402 .

Weilinghoff, A. (2025). Transcribing diverse voices: Using Whisper for ICE Corpora. Proceedings of Interspeech 2025, 3359–3363. https://doi.org/10.21437/Interspeech.2025-1980

Multimodal corpus of AI-generated images: Investigating linguistic bias in English-based visual models

FULL PAPER

Anna Marklová (Charles University, Czech Republic) & Renate Delucchi Danhier (TU Dortmund, Germany)

AI visual art is English-based. This is because text-to-image AI models are trained primarily on data in English and therefore reflect the linguistic and cultural biases embedded in that training. While these systems appear to 'understand' prompts in multiple languages, they typically translate all input into English before generating images (or they fail and generate pictures unrelated to the prompt). Models that disclose their internal prompt reformulations confirm this process. Moreover, a simple input like "A clown is entertaining a queen" may be internally rewritten as "A portrayal of a humorous entertainer amusing a public figure with a crown" to avoid sensitive words.
This study examines the implications of this linguistic mediation for visual output. We present a multimodal corpus of 4,000 AI-generated images paired with their corresponding

external and internal prompts (similar to the Quick, Draw! dataset). The corpus was created by prompting FLUX and DALL·E 3 with 20 English sentences (10 active, 10 passive) describing agent-patient interactions, each prompted 100 times. For each generated image, we recorded the model's internal prompt derived from the initial input and the spatial position of the agent relative to the patient.

Using quantitative corpus-analytic methods, we investigated whether spatial asymmetries known from psycholinguistic and visual cognition research - such as the tendency in left-to-right writing systems to position agents or first-mentioned figures disproportionally on the left (Dobel et al., 2007; Jahn et al., 2007; Maass et al., 2009) are replicated in AI-generated imagery. Results showed that FLUX mirrored human leftward bias in active sentences, but in passive constructions both AI-models positioned the agent on the right significantly more often than humans do.

Our findings suggest that AI image generators not only replicate but amplify Western (specifically English) spatial biases, potentially diminishing the diversity of visual representation shaped by other languages and cultures. This has consequences for the visual landscape around us: AI-generated imagery is increasingly permeating our visual environment, diminishing the diversity in human composition of pictures.

Conceptually, this work demonstrates how multimodal corpora, combining linguistic prompts and generated images, can extend corpus-based research into the domain of AI-mediated visual communication. By integrating quantitative corpus methods with multimodal and cognitive frameworks, this study contributes to the ongoing discussion about the impact of AI on language use and representation.

References

Dobel, Christian, Gil Diesendruck, and Jens Bölte. 2007. "How Writing System and Age Influence Spatial Representations of Actions: A Developmental, Cross-Linguistic Study". Psychological Science 18 (6): 487–491. https://doi.org/10.1111/j.1467-9280.2007.01926.x .

Google Creative Lab. (n.d.). The Quick, Draw! Dataset [Data set]. GitHub. https://github.com/googlecreativelab/quickdraw-dataset

Jahn, Georg & Knauff, Markus & Johnson-Laird, Phil. (2008). Preferred mental models in reasoning about spatial relations. Memory & cognition. 35. 2075-87. 10.3758/BF03192939.

Maass, A., Suitner, C., Favaretto, X., & Cignacchi, M. (2009). Groups in space: Stereotypes and the spatial agency bias. Exp. Soc. Psychol., 45, 496–504. https://doi.org/10.1016/j.jesp.2009.01.004

Why grammatical variation is not necessarily short-lived FULL PAPER

Benedikt Szmrecsanyi (KU Leuven, Belgium)

This talk is an exercise in corpus-based psycholinguistics (see Ma et al. 2025). In the grammar of English, we observe a number of "stable", rather long-lived syntactic alternations, such as the alternation between the ditransitive and prepositional dative (I gave Tom a present vs. I gave a present to Tom), or alternation between particle-object and object-particle order (I looked up the word vs. I looked the word up). To many theorists, the longevity of such variation phenomena, and thus the optionality between alternate ways of saying the same thing (Labov 1972: 188) that they afford, is a bit mystifying. The reason is that this kind of variation is at first glance incompatible with the foundational notion in cognitive linguistics, construction grammar, and functional linguistics that optionality is dysfunctional and thus abnormal. When it occurs, it must – or so many people assume – be suboptimal and short-lived. After all, well-known axioms like the Principle of Isomorphism (Haiman 1980) or the Principle of No Synonymy (Goldberg 1995) imply that language as a complex adaptive system is designed, as it were, to find functionally different niches for particular form-function mappings. If multiple forms are associated with the same meaning or the same grammatical function, it is often predicted that this suboptimal optionality is transitional until languages sort themselves out, so to speak (see De Smet et al. 2018 for critical discussion).

Against this backdrop, we endeavor to test whether grammatical alternations are in a measurable way suboptimal (or: difficult and inconvenient) for language users. We specifically use a corpus-based psycholinguistics research design with a variationist twist and analyze SWITCHBOARD (Godfrey, Holliman & McDaniel 1992), a corpus of conversational spoken American English. We ask if and how grammatical optionality correlates with two symptoms of production difficulty, namely filled pauses (um and uh) and unfilled pauses (speech planning time). Our dataset covers 108,487 conversational turns in SWITCHBOARD, 22 grammatical alternation types yielding 57,032 optionality contexts, 589,124 unfilled pauses and 43,801 filled pauses.

Mixed-effects linear regression analysis shows that overall, optionality contexts do not attract dysfluencies – regardless of how many language-internal probabilistic constraints are in operation, or how many variants there are to choose from. In other words, grammatical optionality does not trigger production difficulties. The interpretation is that there is therefore no production-driven evolutionary pressure to eliminate form-function asymmetry. As we will argue, any additional cognitive inefficiency introduced by having to choose between grammatical alternatives is likely to be offset by a number of compensatory benefits (including adjusting explicitness, managing information density, communicating efficiently, establishing Easy First order, achieving rhythmic well-formedness, domain minimization,

and stalling for planning time).
The take-home message is that grammatical optionality can be long-lived because it is neither difficult and suboptimal, nor dysfunctional. We conclude by discussing a strong interpretation of our results, viz. the Principle of Optionality: "Languages and language users favor the availability of different ways of saying the same thing".

References

De Smet, Hendrik, Frauke D'hoedt, Lauren Fonteyn & Kristel Van Goethem. 2018. The changing functions of competing forms: Attraction and differentiation. Cognitive Linguistics 29(2). 197–234. https://doi.org/10.1515/cog-2016-0025 .

Godfrey, J.J., E.C. Holliman & J. McDaniel. 1992. SWITCHBOARD: telephone speech corpus for research and development. [Proceedings] ICASSP-93.

Goldberg, Adele E. 1995. Constructions: A construction grammar approach to argument structure. Chicago: University of Chicago Press.

Haiman, John. 1980. The iconicity of grammar: Isomorphism and motivation. Language 56(3). 515–540. https://doi.org/10.2307/414448 .

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

Ma, Ruiming, Thomas Van Hoey & Benedikt Szmrecsanyi. 2025. Isomorphism-inspired theorising about optionality and variation: no empirical support from English grammar. English Language and Linguistics 1–21. https://doi.org/10.1017/S1360674325000097 .

Investigating the use of the conjunction and in English prenominal adjective sequences

FULL PAPER

Arne Lohmann & Sarah Clart (Universität Leipzig, Germany)

This study addresses the question of which variables underlie the choice to realize or omit and in English prenominal adjective sequences, see e.g., (1) and (2) below. We report the results of a multifactorial corpus-linguistic analysis testing the claim that this case of variation is driven by both structural-semantic as well as language processing constraints.
(1) A big red ribbon
(2) A big and red ribbon
(3) A (peaceful (political life)).
In the research literature, the use of the coordinating conjunction and is commonly viewed as marking a structural-semantic distinction. Sequences with and are viewed as paratactic constructions in which both adjectives make an "independent attribution of qualities" and that are "of the same kind", typically belonging to the same functional class (Vandelanotte 2002). In contrast, sequences without coordinator or other separator are viewed as instantiating a hypotactic relationship, as illustrated by the bracketing structure in (3), and typically involve adjectives of different functional classes. However, adjective sequences without a coordinator instantiating a paratactic relationship are commonly used, e.g., example (1)

above. We put forth the argument that the choice for/against and is not a purely structural-semantic one, but is also driven by language processing constraints. In particular, we test Rohdenburg's complexity principle, that "in the case of more or less explicit grammatical options, the more explicit one(s) will tend to be favored in cognitively more complex environments." (Rohdenburg 1996: 151). Coordination with and is the more explicit variant and should therefore be preferred in cognitively demanding contexts.

We carried out a multifactorial, corpus-linguistic analysis based on two samples extracted from the Corpus of Contemporary American English (Davies 2014). The samples are a token sample containing 1,000 datapoints of each variant and a type sample that contains all prenominal adjective sequences that occur with and without and in the corpus (n>12,000). As mentioned above, a prediction derived from the structural-semantic account is that the two adjectives combined with and are of the same kind. We tested this claim by coding the semantic classes of the adjectives, as well as measuring their semantic similarity via word embeddings (Günther et al. 2015). Regarding the language processing perspective, a number of variables commonly attributed to impact lexical access were coded, namely length, frequency, concreteness, as well as the semantic and phonological neighborhood density of both adjectives, employing data from the English Lexicon Project (Balota et al. 2007).

A random forest analysis was conducted, showing that variables from both perspectives are statistically significant predictors. Semantically similar adjectives favor sequencing with and, in line with the view of coordination of "the same kind". With regard to the processing-related variables we find that characteristics typically associated with ease/difficulty of retrieval from the mental lexicon yield substantial effects. For example, in case adjectives coordinated are long and/or infrequent, the probability to overtly realize and increases. This is in line with the predictions of the complexity principle. Overall, the results support the interpretation that the phenomenon investigated is co-determined by both semantic-structural considerations as well as processing constraints.

References

Balota, D.A., Yap, M.J., Cortese, M.J., Hutchison, K.A., Kessler, B., Loftis, B., Neely, J.H., Nelson, D.L., Simpson, G.B., & Treiman, R. (2007). The English Lexicon Project. Behavior Research Methods, 39 (3), 445-459.

Davies, Mark. 2014. The Corpus of Contemporary American English: 450 million words, 1990–2012 [Full-Text Corpus Data, offline version of 2014].

Günther, F., Dudschig, C., & Kaup, B. (2015). LSAfun – An R package for computations based on Latent Semantic Analysis. Behavior Research Methods, 47 (4), 930-944.

Rohdenburg, Günter. (1996). Cognitive complexity and increased grammatical explicitness in English. Cognitive Linguistics 7(2), 149-182.

Vandelanotte, Lieven. (2002). Prenominal adjectives in English: Structures and ordering. Folia Linguistica 36 (3-4), 219-259.

Changes in the tense-time relation of 19th and 20th century American English: a study on isomorphism

FULL PAPER

Eva Berlage (Universität Hamburg, Germany)

According to Hawkins (1986), present-day English can be classified as a loose-fit language, which, in comparison to German, typically has a 1:2 or even a 1: many relationship between form and function (see also Steger/Schneider 2012; Berg 2014). In this paper, I want to examine whether we get a decrease in isomorphism (i.e. a 1:1 correlation between form and meaning or form and function; see e.g. Haiman 1980; Givón 1991) in contexts that involve a high degree of redundancy. Contextual redundancy is here defined in the sense that a concept is coded twice, for one via explicit structural markers (e.g. adverbs, conjunctions) and for another via the tense employed (e.g. the present perfect, past perfect).

More specifically, this paper advances the following hypothesis: In 19th and 20th century American English (AmE), we find a decrease in the use of the past perfect and the present perfect in contexts that involve a high degree of contextual redundancy. In the present study, a high degree of contextual redundancy arises if specific temporal adverbs expressing the idea of current relevance (already, just, yet, recently) co-occur with the present perfect or if conjunctions like before and after – where they indicate 'past in the past' (before, after) –combine with the past perfect (see also Kleppel et al. 2021; Kövecses 2000: 184-5). Conversely, redundancy is avoided if the simple past is used in contexts of current relevance and anteriority (in the past). This is indicated in examples (1) and (2), which contain the adverb recently and the conjunction after.

(1) Joseph O'Brien recently bought out Wood's photographer's gallery on the Bowery. (COHA, 1890)

(2) Throughout the early evening, after the children went to bed, the surface of the shell of him was not broken at all. (COHA, 1921)

While previous research has attested to a general decline of the present perfect in AmE from 1750-1800 onwards (see Elsness 1997; 2009) and of the past perfect in the 20th century (see Bowie et al. 2013; Yao/Collins 2013), no one has as yet investigated how these tenses develop in contexts that involve a high degree of contextual redundancy. In my study, I will compare how the use of the simple past develops in high-redundancy contexts.

The corpus employed in the present study is the Corpus of Historical American English (COHA), which spans the time period from 1820-2019. By selecting 500 random tokens (retrieved via the set of temporal adverbs and conjunctions defined above) over five 40-year time periods in the COHA, I will answer the question of whether we see a decrease in the use of the periphrastic present perfect and past perfect over time. If this is the case, I will conclude that isomorphism in the tense system of AmE is on the decline in high-redundancy contexts.

References

Berg, Thomas. 2014. 'Boundary permeability - A parameter for linguistic typology.' Linguistic Typology 18: 489–531.

Bowie, Jill, Sean Wallis and Bas Aarts. 2013. 'The perfect in spoken British English.' In: Bas Aarts, Joanne Close, Geoffrey Leech and Sean Wallis (eds.), The Verb Phrase in English. Investigating Recent Language Change with Corpora. Cambridge: Cambridge University Press. 318–52.

Elsness, John. 1997. The Perfect and Preterite in Contemporary and Earlier English. Berlin: Mouton de Gruyter.

Elsness, Johan. 2009. 'The present perfect and the preterite.' In: Günter Rohdenburg and Julia Schlüter (eds.), One Language, Two Grammars? Differences between British and American English. Cambridge: Cambridge University Press. 228-45.

Givón, Talmy. 1991. 'Isomorphism in the grammatical code: Cognitive and biological considerations.' Studies in Language 15: 85-114.

Haiman, John. 1980. 'The iconicity of grammar: Isomorphism and motivation.' Language 56(3): 515-40.

Hawkins, John. 1986. A Comparative Typology of English and German. Unifying the Contrasts. London/Sydney: Croom Helm.

Kleppel, Stephanie, Matthias Eitelmann and Britta Mondorf. 2021. 'British-American contrasts in the use of the perfect: Negotiation ambiguity versus redundancy?' Zeitschrift für Anglistik und Amerikanistik 69: 291-319.

Kövecses, Zoltán. 2000: American English. An Introduction. Peterborough, Ontario: Broadview Press.

Steger, Maria and Edgar W. Schneider. 2012. 'Complexity as a function of iconicity: The case of complement constructions in new Englishes.' In: Bernd Kortmann and Benedikt Smrecsanyi (eds.), Linguistic Complexity: Second Language Acquisition, Indigenization, Contact. Berlin: Mouton De Gruyter. 156-91.

Yao, X., and P. Collins. 2013. 'Recent change in non-present perfect constructions in British and American English.' Corpora 8: 115–35.

Exploring functional differentiation of near-synonymous pragmatic markers using historical corpora

FULL PAPER

Nicole Benker (Ludwig-Maximilians-Universität München, Germany)

The present study constitutes a corpus pragmatic investigation into the diachronic development of several pragmatic markers from the TRUTH-domain. Previous work on TRUTH-markers, such as honestly and frankly, has found that they tend to develop meanings related to mitigation of face-threats, signaling counterexpectation and concession (Edwards & Fasulo, 2006; Keizer, 2018; Lenker, 2007, 2010; Tseronis, 2011).
The aim of the present study is to investigate the formal and functional development of a subgroup of these markers, which are all derived from clauses that include the lexical

items truth and tell, namely (if the) truth be told, to tell (you) the truth and truth to tell, and which, according to the Oxford English Dictionary (To tell truth, in Tell, v., P.2), are synonymous expressions. Even though it has been shown that functional near-synonymy can be stable over long periods of time (e.g., Torres Cacoullos & Walker, 2009), the idea that near-synonymy must lead to either obsoletion of one form or functional differentiation of all forms is still dominant (De Smet et. al., 2017, Lorenz, 2025). To test whether these formally and functionally similar expressions show signs of functional differentiation, 2572 attestations from the Corpus of Historical American English (data from 1820–2019, Davies, 2008–-b), and 2887 from the Corpus of Contemporary American English (data from 1990–2019, Davies, 2008–-a) were examined.

To tell you the truth, to tell the truth and truth to tell are all attested throughout the entire timeframe covered by both corpora but are characterized by fluctuations in frequency: To tell you the truth is the dominant marker in the early 19th century data but starts to seemingly lose ground to to tell the truth up until 1900, when this development reverses and to tell you the truth becomes dominant again. However, starting in the 1980s, truth be told emerges and becomes the dominant variant, which gives the impression that the expressions are competing for conceptual space.

Contrary to this impression, the investigated markers do not show clear functional differentiation but rather fulfil similar functions to the previously established functions of TRUTH-markers, i.e., counterexpectation and mitigation of face threats. However, they do occur in different genres: To tell the truth and truth to tell mostly occur in fiction writing, while truth be told is dominant in both fiction and TV/movies and to tell you the truth in spoken language and TV/movies. Thus, the aforementioned competition is not based on the development of unique functions but rather use in different genres.

References

Davies, M. (2008–-a). Corpus of contemporary American English (`https://www.english-corpora.org/coca/`)

Davies, M. (2008–-b). Corpus of historical American English (`https://www.english-corpora.org/coha/`)

De Smet, H., D'hoedt, F., Fonteyn, L. & Goethem, K. (2018). The changing functions of competing forms: Attraction and differentiation. Cognitive Linguistics, 29(2), 197–234.

Edwards, D., & Fasulo, A. (2006). "To be honest": Sequential uses of honesty phrases in talk-in-interaction. Research on Language and Social Interaction, 39(4), 343–376.

Keizer, E. (2018). Interpersonal adverbs in FDG - The case of frankly. In E. Keizer & H. Olbertz (Eds.), Recent developments in Functional Discourse Grammar. John Benjamins.

Lenker, U. (2007). Soþlice, forsoothe, truly – communicative principles and invited inferences in the history of truthintensifying adverbs in English. In S. Fitzmaurice & I. Taavitsainen (Eds.), Methods in Historical Pragmatics (pp. 81–105). De Gruyter.

Lenker, U. (2010). Argument and rhetoric: Adverbial connectors in the history of English. De Gruyter.

Lorenz, D. (2024). Potential grammaticalization of epistemic phrases. Functions of Language,

31(3), 262–288.

Oxford University Press. (n.d.). To tell truth, in Tell, v., P.2. In Oxford English dictionary. Retrieved October 31, 2025, from https://doi.org/10.1093/OED/1048814204 .

Schmid, H.-J. (2020). The dynamics of the linguistic system: Usage, conventionalization, and entrenchment. Oxford University Press. Torres Cacoullos, R., & Walker, J. A. (2009). The present of the English future: Grammatical variation and collocations in discourse. Language 85(2). 321–354.

Tseronis, A. (2011). Use and abuse of the strategic function of in fact and frankly when qualifying a standpoint. Pragmatics, 21(3), 473–490.

Combining categorical and multidimensional approaches to situational genre variation in the TRAWL (Tracking Written Learner Language) multilingual corpus

FULL PAPER

Ingrid Kristine Hasund (University of Agder, Norway), Philip Durrant (University of Exeter, UK), Larissa Goulart (Montclair State University, US), Eli-Marie Drange (University of Agder, Norway), Hildegunn Dirdal (University of Oslo, Norway) & Stine Hulleberg Johansen (OsloMet, Norway)

To write successfully at school, students must adapt their texts to diverse genre expectations across different subjects. Learner corpus research (LCR) has long recognised genre as a key variable explaining language variation (Aijmer, 2002; Paquot et al., 2013). Yet, the concept of genre remains contested (Durrant, 2022; Goulart, 2024), and genre labels are often applied inconsistently. For instance, exposition may refer to factual description, explanation, or even argumentation (Durrant, 2022). Such ambiguities challenge both research and pedagogy. Categorical genre typologies (e.g., Nesi & Gardner, 2012; Römer & O'Donnell, 2011) offer familiarity but struggle to capture the hybridity and nuance of real-world student writing. To address the limitations of categorisation, recent work by Biber et al. (2020) and Egbert et al. (2024) reconceptualises genre (defined in their work as registers) as continuous situational variation rather than fixed categories. Using what the authors called situational multidimensional analysis (MDA), this approach describes texts along empirically derived dimensions based on communicative purpose and contextual demands, allowing for both prototypical and hybrid texts to be meaningfully analysed.

This paper presents a novel integration of categorisation and situational MDA to genre analysis, using data from TRAWL - a multilingual corpus of authentic secondary school writing in L1 Norwegian, L2 English, and L3 French, German, and Spanish (Dirdal et al., 2022). While LCR has primarily focused on advanced L2 English writing in one or two genres, little work has examined younger learners in multilingual, multi-genre contexts (cf. Gilquin, 2015; Hasund et al., 2025; Hamann, 2024; Larsson et al., 2021), highlighting a need for studies like

ours. Moreover, we analyse not only learner texts but also the prompts that elicit them, to explore how instructional input influences genre features – addressing another gap in LCR (Melissourgou & Frantzi, 2017).

Drawing on 25 writing prompts (five from each language) and 125 corresponding learner texts (five per prompt) from school year 11, three coders will score situational parameters like This prompt asks the writer to argue, describe, use sources and This text argues, describes, uses sources. In the same process, we record genre labels like essay, keeping numerical situational scores and nominal genre labels clearly distinct. Labels are noted only when explicitly stated in the prompt (e.g., Write an essay) or text (e.g., In this essay, I will…); otherwise, prompts/texts are marked as other. Coder disagreements will be discussed. Scores on situational parameters will be used to produce a quantitative multidimensional profile. Our RQs are:

- Which main situational dimensions and genre labels occur across prompts and texts in the five languages?
- To what extent do genre expectations in prompts align with genre features in texts?

We anticipate that categorisation will capture broad and familiar genre distinctions (Drange, 2025; Durrant, 2025; Hasund, 2022), while MDA will reveal finer-grained variation (Goulart, 2024; Durrant et al., 2025). We also expect mismatches between prompts and texts, especially in L3 writing, where genre instruction is often less explicit (Hamann, 2024). Our integrated approach offers a more nuanced model of multilingual genre variation for research and pedagogy.

References

Aijmer, K. (2002). Modality in advanced learners' written interlanguage. In S. Granger, J. Hung & S. Petch-Tyson (Eds.), Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching (pp. 55–76). John Benjamins. https://doi.org/10.1075/lllt.6.07aij

Biber, D., Egbert, J., & Keller, D. (2020). Reconceptualizing register in a continuous situational space. Corpus linguistics and linguistic theory, 16(3), 581-616.

Dirdal, H., Hasund, I. K., Drange, E-M., Vold, E., & Berg, E-M. (2022). Design and construction of the Tracking Written Learner Language (TRAWL) corpus: A longitudinal and multilingual young learner corpus. Nordic Journal of Language Teaching and Learning, 10(2), 115–135.

Drange, E-M. (2025). La visión de América Latina en textos escritos por alumnos de español como lengua extranjera en Noruega. América Latina en Noruega. De Gruyter, 25-46.

Durrant, P. (2022). Studying children's writing development with a corpus. Applied Corpus Linguistics. https://doi.org/10.1016/j.acorp.2022.100026.

Durrant, P. (2025). What can a corpus tell us about school writing? Findings, challenges, and future directions. Applied Corpus Linguistics. doi: https://doi.org/10.1016/j.acorp.2025.100134

Durrant, P., Goulart., L. & Hasund, I. (2025). Evaluating a non-categorical approach to text types in school and university writing. Conference paper, Register and Task Variation in Learner Corpus Research, UCLouvain, Belgium, 7-8 July.

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. doi: https://doi.org/10.1515/cllt-2024-0011

Gilquin, G. (2015). From design to collection of learner corpora. In S. Granger, G. Gilquin & F. Meunier (Eds.), The Cambridge handbook of learner corpus research (pp. 9–34). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.002

Goulart, L., (2024). Variation in University Student Writing: A Communicative Text Type Approach. John Benjamins. doi:10.1075/scl.117.

Hamann, V. (2024). Lexicogrammatical meaning-making in German as a foreign language: an investigation of Norwegian upper secondary school learners´ texts. PhD thesis. University of Agder. https://hdl.handle.net/11250/3135509

Hasund, I. K. (2022). Genres in young learner L2 English writing. Nordic Journal of Language Teaching and Learning 10 (2), 242-271.

Hasund, I.K., Durrant, P., Goulart, L., Dirdal, H. & Drange, E-M. (2025). Developing a multidimensional tool to characterise types of writing tasks across five languages at secondary level. Conference paper, Register and Task Variation in Learner Corpus Research, UCLouvain, Belgium, 7-8 July.

Larsson, T., Paquot, M., & Biber, D. (2021). On the importance of register in learner writing: A multi-dimensional approach. In E. Seoane & D. Biber (Eds.), Corpus based approaches to register variation (pp. 235-258). Benjamins.

Melissourgou, M. N., & Frantzi, K. T. (2017). Genre identification based on SFL principles: The representation of text types and genres in English language teaching material. Corpus Pragmatics, 1(4), 373–392.

Nesi, H., & Gardner, S. (2012). Genres across the disciplines: Student writing in higher education. Cambridge University Press.

Paquot, M., Hasselgård, H., & Ebeling, S.O. (2013). Writer/reader visibility in learner writing across genres: A comparison of the French and Norwegian components of the ICLE and VESPA learner corpora. In S. Granger, G. Gilquin & F. Meunier (Eds.), Twenty Years of Learner Corpus Research: Looking back, Moving ahead (pp. 377–387). Presses universitaires de Louvain.

Römer, U., & O'Donnell, M. B. (2011). From student hard drive to web corpus (part 1): the design, compilation and genre classification of the Michigan Corpus of Upper-level Student Papers (MICUSP). Corpora, 6(2), 159-177. doi: https://doi.org/10.3366/cor.2011.0011

Tracing collocational development in intermediate learners' writing: A truly longitudinal study

FULL PAPER

Rolf Kreyer & Sandra Götz (Marburg University, Germany)

The development of collocational competence is a crucial aspect of advanced language proficiency (Nesselhauf, 2005; Schmitt, 2010). Research has consistently shown that intermediate learners tend to underuse conventional collocations and rely heavily on semantically trans-

parent or non-idiomatic combinations (e.g. Bestgen & Granger, 2014; Durrant & Schmitt, 2009). As proficiency increases, learners generally display greater collocational diversity and appropriateness (Siyanova & Schmitt, 2008; Paquot & Granger, 2012). However, much of this research is typically based on cross-sectional designs (e.g. Granger & Bestgen 2014). Truly longitudinal studies are either restricted in terms of number of informants (e.g. Li et al. 2014) or in terms of time span covered (e.g. Siyanova-Chanturia & Spina 2020). Consequently, we still know relatively little about how individual learners' collocational repertoires develop over extended periods of instructed learning, leaving a gap in our understanding of the development of phraseological competence in learner language.

Against this background, the present study contributes to this underexplored area by examining the collocational development of 92 German learners of English over a four-year period, i.e. from grades 9 through to 12. Drawing on the Marburg Corpus of Intermediate Learner English (MILE; Kreyer, 2015), which comprises multiple writing samples from each learner over a period of 4 years, we will identify and quantify the development of different types of collocations, including lexical, grammatical, and academic collocations. Collocations are identified using both statistical association measures (e.g. t-score, MI) and reference resources such as the Pearson Academic Collocations List (Ackermann & Chen, 2013). To capture developmental patterns, we will employ mixed-effects regression models (Gries, 2021), allowing us to assess the influence of grade level, lexical frequency, task type and learner-specific factors on collocational use. The study seeks to answer the following research question:

How do different types of collocations develop in the written English of German school learners over four years of secondary education?

Preliminary analyses based on the Pearson Academic Collocations List (2,469 items) indicate a steady increase in the use of academic collocations from Grade 9 to Grade 12, showing that growing lexical sophistication correlates with a greater uptake of formulaic academic language (Paquot, 2019; Durrant, 2017). We expect further analyses to reveal possible different growth patterns across collocation types, with lexical and grammatical collocations possibly developing earlier and academic collocations showing increases only later in the data.

By offering a true-longitudinal and multilayered perspective on collocational development in intermediate learners, this study provides new insights into how phraseological competence develops during secondary school.

References

Ackermann, K., & Chen, Y. (2013). The Academic Collocations List: A corpus-driven list of frequent collocations in written academic English. TESOL Quarterly, 47(4), 842-848.

Bestgen, Y., & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. Journal of Second Language Writing, 26, 28-41.

Durrant, P. (2017). Lexical bundles and disciplinary variation in university student writing. Applied Linguistics, 38(2), 165-193.

Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of

collocations? International Review of Applied Linguistics, 47(2), 157-177.

Granger, S. & Bestgen, Y. (2014). The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. International Review of Applied Linguistics in Language Teaching, 52, 229-252.

Gries, S. Th. (2021). (Generalized Linear) Mixed-Effects Modeling: A Learner Corpus Example. Language Learning, 71(3), 757-798.

Kreyer, R. (2015). The Marburg Corpus of Intermediate Learner English (MILE). In M. Callies & S. Götz (eds), Learner Corpora in Language Testing and Assessment (pp. 13-34). Amsterdam: John Benjamins.

Nesselhauf, N. (2005). Collocations in a learner corpus. Amsterdam: John Benjamins.

Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. Second Language Research, 35(1), 121-145.

Paquot, M., & Granger, S. (2012). Formulaic language in learner corpora. Annual Review of Applied Linguistics, 32, 130-149.

Schmitt, N. (2010). Researching vocabulary: A vocabulary research manual. Palgrave Macmillan.

Siyanova, A., & Schmitt, N. (2008). L2 learner production and processing of collocation: A multi-study perspective. Canadian Modern Language Review, 64(3), 429-458.

Siyanova-Chanturia, A. & Spina, S. (2020). Multi-Word Expressions in Second Language Writing: A Large-Scale Longitudinal Learner Corpus Study. Language Learning, 70, 420-463.

Syntactic Complexity Development of Intermediate Learner English: A longitudinal Corpus Based Study

FULL PAPER

Philine Metzger (Philipps-University Marburg, Germany)

Complexity is a topic of wide range, overall separated into various levels, namely lexis (Linnarud, 1986), morphology (Brezina & Pallotti, 2019), syntax (Lee, 2004) and phraseology (Paquot, 2019). Syntactic complexity can be described as variation and sophistication of syntactic units or grammar (Foster & Shekan 1996, Ortega 2003). It is generally assumed that syntactical complexity of a language correlates positively with overall language proficiency and development. (Bulté 2008 & Paquot, Naets, Gries 2021) Lu suggests that complexity can, among others, be measured with 14 parameters of complexity that are generally divided into five overall groups, namely length of production unit, particular structures, sub- and coordinate phrases and sentence complexity ratio (Lu 2010: 479). Previous studies have already analyzed syntactic complexity in learner language, whereas their corpora were much smaller, metadata was missing and it was not compared to qualitative development of learner texts. (Kyle, Crossley, Verspoor, 2021).
The present project therefore poses the following research question:
How does syntactic complexity in written English of intermediate learners of a German high school develop quantitatively and qualitatively between the 9th and 12th grade?
To answer this research question, the “Marburg Corpus of Intermediate Learner English”

(MILE; Kreyer, 2015) enables the opportunity to analyze data of 88 students from 9th to 12th grade. With more than 500,000 words in total, the MILE corpus conducted a truly longitudinal corpus and provides the chance of analyzing a large number of German intermediate learners of English over four years including different metadata such as gender, age and frequency of usage of English. The corpus has been analyzed by extracting the previously mentioned 14 parameters of complexity with the "Tool for Automatic Analysis of Syntactic Sophistication and Complexity" (TAASSC; Kyle, 2016). Additionally to the quantitative results, the study will conduct teacher's ratings to reveal a perspective on the qualitative development as well. Therefore, teachers will grade learner texts in twelve different parameters such as lexis and sentence variation, overall correctness in grammar, spelling and sentence structures.

The method that will be used replicates previous studies connected to syntactic complexity development (Crossley & McNamara, 2012) and follows the "bottom-up" principle, therefore, data will be analyzed in the first step and will be put into a larger context, namely through adding metadata and comparing it to qualitative results, in the second step.

First results of the whole data show an overall increase of quantitative complexity which is variant through the different parameters and needs to be evaluated further in the next step. Finally, the project will result in an individual and multifactorial learner curve and discuss language pedagogical usage. So far, a pilot study has successfully been conducted and showed how the used tools are appropriate for the specific corpus and first insights within the results of the pilot study also showed a qualitative improvement over the course of the years. (Götz-Lehmann, Kettenhofen, Metzger, tba).

References

Brezina, Vaclaw; Pallotti, Gabriele (2019): Morphological Complexity in Written L2 Texts. Second Language Research. 35 (1): 99-119.

Bulté, Bram; Housen, Alex; Pierrad, Michel; Van Daele, Siska (2008): Investigating Lexical Proficiency Development over Time – The Case of Dutch-Speaking Learners of French in Brussels. Journal of French Language Studies 18 (3): 277-298.

Crossley, Scott; McNamera, Danielle (2012): Predicting Second Language Writing Proficiency: The Roles of Cohesion and Linguistic Sophistication. Journal of Research in Reading 35 (2): 115-135.

Foster, P., & Skehan, P. (1996). The influence of planning time on performance in task-based learning. Studies in Second Language Acquisition, 18, 299–234. https://doi.org/10.1017/S0272263100015047

Götz, Sandra; Kettenhofen, Fabian; Metzger, Philine (tba: 2026): Syntactic Complexity Development of Intermediate L2 English: A longitudinal, corpus-based study (working title).

Kreyer, Rolf (2015): The Marburg Corpus of Intermediate Learner English (MILE). In Learner Corpora in Language Testing and Assessment, Marcus Callies & Sandra Götz, eds. Amsterdam: John Benjamins. 13-34.

Kyle, K. (2016): Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication (Doctoral Dissertation).

Lee, Jiyoujng (2004): Syntactic Complexity, Clausal Complexity, and Phrasal Complexity in L2

Writing: The Effects of Task Complexity and Task Closure. The Journal of Asia TEFL. South Korea: 108-124.

Linnarud, Moira (1986): Lexis in Composition: A Performance Analysis of Swedish Learners Written English. Malmö: C.W.K. Gleerup.

Ortega, Lourdes (2003): Syntactic Complexity Measures and their Relationship to L2 Proficiency: A Research Synthesis of College-level L2 Writing. Applied Linguistics 24 (4): 492-518. https://doi.org/10.1093/applin/24.4.492

Paguot, Magali (2019): The Phraseological Dimension in Interlanguage Complexity Research. Second Language Research 35 (1): 121-145.

Paquot, Magali; Naets, Hubert; Gries, Stefan (2021): Using Syntactic Co-occurrences to Trace Phraseological Complexity Development in Learner Writing: Verb + Object Structures in LONGDALE. In: LeBruyn, Bert Simonne Walter; Paquot, Magali (hrsg): Learner Corpus Research Meets Second Language Acquisition. Cambridge: Cambridge University.

ASR and Multimodal AI — WIP [2] E113 • 16:00–17:00

A corpus for studying voice-AI bias and dialectal adaptation in Newcastle English

WIP

Dana Serditova (University of Regensburg, Germany, Heinrich Heine University Düsseldorf, Germany), Kevin Tang (Heinrich Heine University Düsseldorf, Germany) & Jochen Steffens (University of Applied Sciences Düsseldorf, Germany)

Voice technologies often struggle with regional dialects due to training data biased toward mainstream varieties. There is also a gap in research when it comes to dialectal adaptation and communicative strategies that speakers employ when using voice technologies. We present a new corpus designed to investigate how speakers adjust their speech when interacting with voice technologies that do or do not accommodate regional dialects. The dataset comprises interactions between participants and two types of Voice Assistants (VAs): one that recognizes and responds in the speaker's regional dialect, and one that uses only a mainstream variety. We focus on (1) the retention or suppression of salient dialectal features and (2) communicative strategies participants employ, such as convergence and divergence.

The regional focus is on Newcastle English. Aside from being one of the most well-studied dialects (Mearns 2015, Hughes et al. 2013), Newcastle English is widely recognized and closely tied to local identity. Its speakers are also aware of the distinctiveness of their dialect (Beal 2009, Pearce 2017). Moreover, it is one of the most challenging dialects for speech technology (Markl 2022). We aim to collect data from 100 participants, balanced by gender, age, and socioeconomic background.

The corpus is being compiled using the Wizard-of-Oz method, which simulates a fully functioning VA — a necessary approach given that no commercial VA currently supports

Newcastle English. In this setup, experimenters ("wizards") manually control the system's responses to the user's input, making it appear as if the system is operating autonomously. The VA replies were delivered using pre-recorded speech by a native Newcastle English voice actor, in Newcastle English for the dialectal system and in Standard Southern British English for the mainstream system.

In order to control the flow of the interaction, participants are offered to make requests to the system by means of two types of guided prompts: (1) "sociolinguistic" prompts designed to elicit salient phonological and morphosyntactic features, and (2) "semantic" prompts that test whether different situational contexts (a medical emergency scenario, vacation planning, etc.) have an additional effect on the realization of regional linguistic features. Since the participants make requests freely instead of reading out the prompt, the target features are systematically elicited without compromising naturalness.

In this work-in-progress report, we introduce the design and structure of the corpus, present our stimuli and experimental setup, and share initial findings on speech adaptation and participant experiences and perceptions of the system. We also discuss preliminary results on the correlation between the realization of linguistic features and sociolinguistic variables such as age, gender, educational background, and prior experience with or exposure to VAs.

These findings have important implications for understanding how speech technologies may contribute to processes of social and cultural erasure in AI-mediated communication, particularly with respect to the marginalization of non-mainstream identities. To our knowledge, this is the first corpus specifically designed to examine dialectal adaptation and communicative strategies in interactions with voice technologies that (do or do not) accommodate regional dialects.

References

Beal, J. C. (2009). Enregisterment, commodification, and historical context: "Geordie" versus "Sheffieldish". American Speech 84(2). 138–156.

Hughes, A., P. Trudgill & D. Watt. (2013). English accents and dialects: An introduction to social and regional varieties of English in the British Isles. Routledge.

Koenecke, A., A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky & S. Goel. (2020). Racial disparities in automated speech recognition. Proceedings of the National Academy of Sciences 117(14). 7684–7689.

Markl, N. (2022). Language variation and algorithmic bias: understanding algorithmic bias in British English automatic speech recognition. In Proceedings of the 2022 ACM FAccT Conference, 521–534.

Mearns, A. (2015). Tyneside. In R. Hickey (ed.), Researching Northern English, 161–182. John Benjamins Publishing Company.

Pearce, M. (2017). The linguistic landscape of North-East England. Perspectives on Northern Englishes 96. 61–82.

What it's like to compile a multimodal corpus of your own speech – and its implications for AI-generated podcasts

WIP

Christina Sanchez-Stockhammer (Technische Universität Chemnitz, Germany) & Dominic Piazza (Technische Universität Chemnitz, Germany (associated))

The goal of this work-in-progress report is to introduce a new, fully transcribed linguistics podcast as an original resource for corpus linguistic research, to share the unusual stories behind its compilation by the podcast hosts themselves, and (in line with the ICAME47 topic, "A Confluence of Corpus Research in the Age of AI") to discuss the potential of the corpus to serve as the basis for AI-generated episodes.

The podcast Linguistics Behind the Scenes (https://www.mytuc.org/linguisticsBTS) targets general audiences and can be accessed on various podcast platforms. It attempts to instruct while entertaining and avoids the use of linguistic terminology as much as possible. Episode length varies between 20 and 79 minutes. In each of the more than 20 episodes so far, the two hosts have informal dialogic conversations about language, culture and their own linguistic research. In line with Bell & Gibson (2011: 558), this communicative situation represents a "mediated staged performance" of linguistic discourse. The recorded conversations consist of spontaneous language with slight post-editing in the form of cuts.

The LinguisticsBTS Corpus is a multimodal corpus of spoken English, which is open-ended, like the podcast itself. It consists of the published audio files (`https://www.youtube.com/@LinguisticsBTS`) and a compilation of transcripts for all episodes on a single webpage (`https://www.tu-chemnitz.de/phil/english/sections/edling/sciencecommunication/podcast-transcripts.php`). The written corpus currently comprises over 120,000 words of orthographic transcription including episode titles and the names of the two hosts (Christina and Dominic) followed by colons in the beginning of each turn. The transcription software employing artificial intelligence built into Apple's iOS is used to produce rough versions of the transcripts from the audio files, but important amounts of human post-editing are required to mark metalanguage with italics and quotation marks and to ensure precision and proper spelling, e.g. in the frequent instances of code-switching between languages (such as English, German, Spanish, French or Japanese). For better readability (and to serve its main function as support for listeners of the podcast), the transcripts do not conserve all hesitation phenomena and repetitions, but still permit the study of many features of spoken English (e.g. sentence structures and broken-off constructions). With one of the hosts following the standards of British and American English each, it can also serve the comparison of these two varieties and to investigate the register of science podcasts.

Unlike the usual situation in corpus compilation, which makes use of texts produced by other speakers, the LinguisticsBTS Corpus contains the speech of the corpus compilers themselves. This work-in-progress report therefore offers unique insights into what it is like for linguists to reflect their own spoken language use while post-editing long transcripts of their own speech (e.g. with regard to AI hallucinations, decision-making on sentence boundaries and idiosyncratic language use).

With its free online availability, the LinguisticsBTS corpus can also serve as a resource

for training artificial intelligence on typically human features of science podcasts and to investigate similarities and differences to AI-generated podcasts. The closing part of this work-in-progress report provides some first insights into an ongoing case study in this area.

References

Bell, Allan & Andy Gibson. 2011. Staging language: An introduction to the sociolinguistics of performance. Journal of Sociolinguistics 15(5). 555-572.

Corpus Grammar Research — WIP [3] E114 • 16:00–17:00

Testing the bell curve hypothesis: How grammaticalization stages shape register variation in probabilistic grammar

WIP

Claudia Thorwarth (Leipzig University, Germany)

This work-in-progress paper presents preliminary findings from a corpus-based investigation into how grammaticalization stages systematically influence register variation in probabilistic grammar. Building on recent work showing that grammatical alternations exhibit varying degrees of register stability (Engel & Szmrecsanyi 2022; Bartels & Szmrecsanyi 2024), I propose a bell curve hypothesis on the relation between grammaticalization and probabilistic grammar: alternations at intermediate grammaticalization stages show maximum register divergence, while both early and late stages exhibit greater stability. This is tested using 12,000 concordances on the future temporal reference and deontic modal variations from the Corpus of Contemporary American English (COCA, Davies 2008-) across six registers.

The analysis examines multiple alternation phenomena positioned at different points along the grammaticalization continuum. Future temporal reference (will vs. be going to) represents an intermediate stage, with the incoming be going to not being an established marker of futurity until the middle of the 19th century (Mair 2004: 129) and still showing vigorous change both qualitatively and quantitatively (Mair 1997: 1541). Deontic modality (must vs. have (got) to) involves a more advanced alternation, with the two forms competing since at least Middle English, and the incoming have (got) to even overtaking must during the late 20th century based on quantitative accounts (Tagliamonte & Smith 2006: 347-348; Leech et al. 2009: 73). Using Variation-based Distance and Similarity Modelling (VADIS, Szmrecsanyi et al. 2019), mixed-effects models and random forests were fitted to measure register divergence through three metrics: predictor significance patterns, effect sizes, and constraint hierarchies.

Preliminary results support the bell curve hypothesis. Future temporal reference shows high register instability (mean distance score: 0.55), with especially predictor significance varying dramatically – the factor animacy, for instance, ranges from highly significant in

academic writing to non-significant in spoken discourse. Deontic modality exhibits greater stability (mean distance score: 0.39). To further test the bell curve model's boundaries, two additional phenomena hypothesized to occupy the extremes of the grammaticalization continuum will be analyzed. The get-passive, still emerging and comparatively still low-frequent in contemporary English, potentially represents a very early grammaticalization stage. Conversely, the s-genitive, which has undergone centuries of grammaticalization and shows signs of late-stage specialization, may exemplify an advanced stage. These analyses will test whether phenomena at the edges of the continuum indeed show the predicted lower register divergence, thus providing further proof to the bell curve model.

This work contributes to understanding the relationship between diachronic change and synchronic variation. The bell curve pattern suggests that grammaticalization creates predictable trajectories of register variation: emerging forms show relative stability as they establish their functional niche, maximum divergence occurs during active competition and specialization across registers, and eventual re-stabilization emerges as forms develop specialized functions. I propose that grammaticalization leaves "uneven traces" across usage domains, with register divergence patterns influenced by the stage of grammatical development.

References

Bartels, B. & Szmrecsanyi, B. (2024). Future temporal reference in spoken world Englishes. World Englishes, 1–17.

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

Engel, A. & Szmrecsanyi, B. (2022). Variable grammars are variable across registers. Language Variation and Change 34, 355–378.

Mair, Christian. (1997). The spread of the going-to-future in written English: A corpus-based investigation into language change in progress. In Raymond Hickey & Stanislav Puppel (Eds.), Language history and linguistic modelling (pp. 1537–1543). Mouton de Gruyter.

Mair, Christian. (2004). Corpus linguistics and grammaticalisation theory: Statistics, frequencies, and beyond. In Hans Lindquist & Christian Mair (Eds.), Corpus Approaches to Grammaticalization in English (pp. 121–150). John Benjamins Publishing Company.

Szmrecsanyi, B., Grafmiller, J. & Rosseel, L. (2019). Variation-Based Distance and Similarity Modeling. Frontiers in Artificial Intelligence 2.

Tagliamonte, Sali & Jennifer Smith (2006). Layering, competition and a twist of fate: Deontic modality in dialects of English, Diachronica, 23(2): 341-380.

A fresh look at infinitive variation after help in 19th–21st century American and British English

WIP

Richard Zimmermann (University of Manchester, United Kingdom)

The verb help shows variation in the form of its non-finite complement, which can be headed either by a to-infinitive or a bare infinitive, as illustrated in (1) and (2) respectively.
(1) It helped to lower the cost of imports (BNC, A6F 246)
(2) It helped improve project design (BNC, J3L 217)
The variation is the result of a change. While the non-finite marker to was virtually always present until the early 19th century, the bare infinitive has now become dominant.
Building on previous research (McEnery & Xiao 2005, Lohmannn 2011, Levshina 2022), this study builds a new, larger dataset based on American and British English fiction and non-fiction texts collected from c. 800 million words in the corpora Evans (TCP 2014), COHA (Davies 2010), COCA (Davies 2008), CLMET (Diller et al. 2011), HUM19UK (Walker et al.), BNC (BNC Consortium 2007), and BNC2014 (Love et al. 2017). Mixed-effects logistic regression models are constructed to investigate the following research questions.
First, the time course of the development will be determined in far greater detail than previously. It will be possible to ascertain the rate of change as well as the likely time and variety of the actuation of the change, and its state in Present-Day English. Initial results suggest the origin of the bare infinitive in American English, and catch-up in innovation in British English in the middle of the 20th century. Secondly, a model will be built for the hitherto under-investigated variant of -ing form complements, as in (3). Polarity will be taken into account (note not in (3)).
(3) I could not help acknowledging the truth of his assertion (COHA, Fic, 1839)
Preliminary results suggest dominance of -ing complements in British English, as illustrated in Figure 1. This could have impeded the adoption of the bare infinitive in that variety.
Figure 1: Plot of -ing participle vs. bare and full infinitives after help, colour-coded for Variety, data points represent texts, point size proportional to number of examples per text, horror aequi (to help) removed, cannot help but do patterns removed, regression lines from mixed-effects logistic regression predicting infinitive from year, variety, and their interaction, including random text intercepts.
Lastly, the change will be scrutinised for hallmarks of grammaticalisation (e.g., Heine 2003). This is motivated by the fact that bare infinitives usually occur in functional structures (e.g., after modals, let's, AcI verbs). Results supporting grammaticalisation should show an increase in the relative frequency of help over time, and in the proportion of (an expected 5,000) it subjects that are dummies, as in (4).
(4) Sometimes [expletive it] helps both parties for them [associate to have the United States as the mediator] (from McEnery & Xiao 2005: 184: ex. (22b)) The study's significance lies in providing unprecedented detail on help's historical trajectory through robust statistical modelling.

References

BNC Consortium (2007) The British National Corpus, XML Edition. Oxford Text Archive. http://hdl.handle.net/20.500.14106/2554 (last accessed 13/11/2025).

Davies, Mark (2008) The Corpus of Contemporary American English (COCA). https://www.english-corpora.org/coca/ (last accessed 13/11/2025).

Davies, Mark (2010) The Corpus of Historical American English (COHA). https://www.english-corpora.org/coha/ (last accessed 13/11/2025).

Diller, Hans-Jürgen, De Smet, Hendrik, Tyrkkö, Jukka (2011) 'A European database of descriptors of English electronic texts.' The European English Messenger 19, 21–35.

Heine, Bernd (2003) 'Grammaticalization.' In: Joseph Brian D. & Janda, Richard D. (eds.) The Handbook of Historical Linguistics. Malden, MA: Blackwell, 573–601.

Levshina, Natalia (2022) 'Comparing Bayesian and Frequentist Models of Language Variation: The Case of Help + (to-)Infinitive.' In: Schützler, Ole & Julia Schlüter (eds.) Data and Methods in Corpus Linguistics. Cambridge: Cambridge University Press, 224–258.

Lohmannn, Arne (2011) 'Help vs. help to: A multifactorial, mixed effects account of infinitive marker omission.' English Language and Linguistics 15.3, 499–521.

Love, Robin, Dembry, Chris, Hardie, Andrew, McEnery, Tony and Piao, Nelan (2017) 'The spoken BNC2014: Designing and building a spoken Corpus of everyday conversations.' International Journal of Corpus Linguistics 22.3, 319–44.

McEnery, Anthony & Zhonghua Xiao (2005) 'HELP or HELP to: What Do Corpora Have to Say?' English Studies 86.2, 161–187.

TCP (Text Creation Partnership) (2014) Evans Early American Imprints Online. https://quod.lib.umich.edu/e/evansdemo/ (last accessed 13/11/2025).

Walker, Brian, McIntyre, David, Land, Elliott, Price, Hazel, & Burke, Michael (2019) HUM19: the Huddersfield-Utrecht-Uppsala-Middelburg Corpus of 19th Century British and Irish Fiction. https://www.uu.se/en/department/english/research/english-linguistics/electronic-resource-projects/hum19-huddersfield-utrecht-uppsala-middelburg-corpus-of-19th-century-british-and-irish-fiction (last accessed 13/11/2025).

Variation and change in the network of argument structure constructions (ASCs): a multivariate study of the alternation between the 'way'-construction and some of its variants in Early Modern to Present-Day English

WIP

Jimena Manuela Jiménez Real (University of Helsinki, Finland)

The development of the 'way'-construction ("Sam joked his way out of the meeting") has been extensively studied by linguists (e.g., Israel, 1996; Fanego, 2018; Perek, 2018). Estimated to have arisen in the 17th century, it has evolved through a process of semantic extension that has resulted in an increasing productivity of its verb slot (Perek, 2018).
In recent times, constructions have begun to attract the attention of historical sociolinguists

(e.g., Säily et al., 2025). The benefits of studying language by combining Construction Grammar (CxG) with historical sociolinguistics have been remarked upon, for instance, by Hilpert (2017). Notably, studying CxG from the perspective of historical sociolinguistics can provide information on the social factors that have influenced the emergence and development of constructions. The first investigations of the 'way'-construction in this dual framework have shown that the construction has been used with varying productivity over time and across social groups (Perek et al., 2024; Jiménez Real, 2025). One possible explanation for the social variation is that groups who use the construction less innovatively instead resort to parallel constructions, or variants.

In my doctoral research, I intend to explore variation and change in the alternation between the 'way'-construction and two other Argument Structure Constructions (ASCs) (see e.g. Perek, 2015; Goldberg, 2019): the Intransitive Motion Construction (IMC, "Skiers whooshed down the slopes", studied by Fanego, 2018) and the reflexive ("He drank himself to oblivion", investigated by Mondorf, 2011). My methods include corpus linguistics, inferential statistics, and using large language models (LLMs) to annotate social metadata. My data will consist of various corpora that enable the study of variation across social groups, including gender variation: the Corpora of Early English Correspondence (CEEC, 1400-1800), Early English Books Online (EEBO, 1473-1700) and the Corpus of Historical American English (COHA, 1810-2009). The goals of my research include (1) assessing how the distribution of both variants of each alternation has changed from Early Modern English to Present-Day English and (2) examining which language-internal and external factors have impacted the choice between the two variants.

At this point of my research, I have begun to examine the interaction between the 'way'-construction and the Intransitive Motion Construction (IMC) in EEBO. The language-internal factors that I plan to analyze include the meaning of the construction ('means' or 'manner'), priming, the type of path (in terms of the preposition or adverb that heads the construction's directional phrase, concrete/abstract path, metaphoric/literal path) and the animacy of the subject. The language-external factors that I intend to analyze are gender, genre and profession. In order to annotate social metadata in EEBO I'm considering the option of using LLMs (Säily et al., 2025). My presentation will discuss my findings from EEBO regarding how language-internal factors and language-external factors interacted to motivate speakers' choice between the 'way'-construction and the IMC in the early stages of development of the 'way'-construction.

References

Fanego, T. (2018). A construction of independent means: the history of the Way construction revisited. English Language and Linguistics, 23(3), 671–699. https://doi.org/10.1017/S1360674318000059

Goldberg, A. E. (2019). Explain me this: creativity, competition, and the partial productivity of constructions. Princeton University Press. https://doi.org/10.1515/9780691183954

Hilpert, M. (2017). From mutual challenges to mutual benefits: Historical sociolinguistics and construction grammar. In T. Säily, A. Nurmi, M. Palander-Collin & A. Auer. (eds.), Exploring

Future Paths for Historical Sociolinguistics (pp. 217–237). John Benjamins Publishing Company. https://doi.org/10.1075/ahs.7.09hil

Israel, M. (1996). The "way" constructions grow. In Goldberg, A. (ed.), Conceptual Structure, Discourse and Language, (pp. 217-230). CSLI.

Jiménez Real, J. M. (2025). Who led the semantic extension of the way-construction? A corpus-based study of variation and change in the way-construction in the 17th-19th centuries [Master's thesis, University of Helsinki]. Helda. http://hdl.handle.net/10138/594920

Mondorf, B. (2011). Variation and change in English resultative constructions. Language Variation and Change, 22(3), 397–421. https://doi.org/10.1017/S0954394510000165

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.https://doi.org/10.1515/cllt-2016-0014

Perek F., Säily T. & Suomela J. (2024, August). Historical sociolinguistics meets constructional change: Gender and the way-construction in the Corpus of Historical American English [presentation]. 57th Annual Meeting of the Societas Linguistica Europaea (SLE 2024), Helsinki, Finland.

Säily, T., Perek, F. & Suomela, J. (2025). Variation and change in the productivity of BE going to V in the Corpus of Historical American English, 1810–2009. In T. Säily & T. Vartiainen (eds.), English Language and Linguistics 29(2), 362–388. Special issue, Constructional approaches to creativity and productivity in English.

Säily, T., Suomela, J., Perek, F., Jiménez Real, J., & Vartiainen, T. (2025). Using large language models to enrich corpus metadata: The case of novels in COHA. Paper presented at the 46th Annual Conference of the International Computer Archive of Modern and Medieval English (ICAME 46), Vilnius, Lithuania, June 2025. https://tanjasaily.fi/talks/icame46_saily_et_al_2025.pdf

Historical Linguistics — WIP [3] E313 • 16:00–17:00

Middle English Thesaurus of the Learned Professions (METtLe-Pro)

WIP

Johanna Vogelsanger & Anthony Harris (University of Zurich, Switzerland)

In this work-in-progress report, we would like to present our current project, *Waxing and Waning Words: Lexical Variation and Change in Middle English* (WAW-ME): an online database of Middle English vocabulary of the learned professions (METtLe-Pro), i.e., that pertaining to four domains – law, medicine, religion, and education. The project will run for four years and the project's overall research questions concern diachronic, diatopic, and diaphasic variation and change across the Middle English period, as well as how native vs. borrowed lexemes are distributed within the semantic hierarchies of the four domains, whether the four domains show significant differences in these developments, and whether all of these processes work towards lexical standardisation in the four domains.
The lexical items for the database are identified by consulting the available lexicographical

resources for Middle English: the *Historical Thesaurus of English* (HTE), the *Middle English Dictionary* (MED), and the *Oxford English Dictionary* (OED). Additionally, the data is cross-referenced in the Old English resources, i.e., the *Dictionary of Old English* (DOE), Bosworth-Toller's *Anglo-Saxon Dictionary* (BT), and the *Thesaurus of Old English* (TOE). Due to the influence of Latin and French during the Middle English period, we also consult the relevant dictionaries for these languages, such as the *Dictionary of Medieval Latin from British Sources* (DMLBS), the *Anglo-Norman Dictionary*, and the *Dictionnaire du Moyen Français* (DMF).
The lexical items thus collected and included in the database are then categorised according to broad semantic frames. These frames vary slightly according to domain but are roughly compatible and can be summarised as follows: Agents and Participants, Institutions, Actions and Processes, Objects, States and Conditions, Specific Times, and Specific Places. The categorisation is based on the dictionary definitions. Within each broad frame, the lexical items are further organised into several hierarchical levels according to their specificity.
Once the database has been compiled, it will serve as a basis for further, corpus-based investigations, which will serve to answer the project's main research questions. For the ICAME47 conference, however, we would like to present an overview and summary of the database and the work that has gone into it: its contents, the public-facing presentation of it, challenges pertaining to data collection and classification. By the time of the conference, we expect to prepare a sample of a few hundred words per domain entered into the database and a working website for demonstration purposes.

References

BT = Bosworth, J. and T. Northcote Toller. (1898-1921). An Anglo-Saxon dictionary, based on the manuscript collections of the late Joseph Bosworth. Clarendon Press. Online edition ed. Ondrej Tichy, Martin Rocek et al. Charles University Prague. https://bosworthtoller.com/

DMF = Dictionnaire du Moyen Français. (2023). ATILF - CNRS & Université de Lorraine. http://www.atilf.fr/dmf

DMLBS = Dictionary of Medieval Latin from British Sources online. (2012-). Based on print edition, ed. R. K. Ashdowne, D. R. Howlett, & R. E. Latham. British Academy/Oxford University Press. https://clt.brepolis.net/dmlbs/Default.aspx

DOE = Dictionary of Old English: A to Le online. (2024). Ed. Angus Cameron, Ashley Crandell Amos, Antonette diPaolo Healey et al. Toronto: Dictionary of Old English Project.

HTE = Kay, Christian, Marc Alexander, Fraser Dallachy, Jane Roberts, Michael Samuels, and Irené Wotherspoon (eds.). (2025). The Historical Thesaurus of English (2nd edn., version 5.0). University of Glasgow. https://ht.ac.uk/.

MED = Middle English Dictionary. (1952-2001). Ed. Robert E. Lewis, et al. University of Michigan Press. Online edition in Middle English Compendium. Ed. Frances McSparran, et al.. (2000-2018). University of Michigan Library. http://quod.lib.umich.edu/m/middle-english-dictionary/

OED = Oxford English Dictionary online. (2025). Oxford University Press. https://www.oed.com

TOE = Roberts, Jane and Christian Kay with Lynne Grundy. (2017). A Thesaurus of Old English. University of Glasgow. http://oldenglishthesaurus.arts.gla.ac.uk/

The Fairest in the Land: Verbatim Transcriptions and Paralinguistic Features in the Mirror of Parliament Corpus (1828-1841)

WIP

Marc Alexander (University of Glasgow, United Kingdom)

This short paper presents work in progress on a new corpus of The Mirror of Parliament, a weekly publication reporting parliamentary debates between 1828 and 1841. The Mirror has received very little scholarly attention but offers a highly unusual resource: verbatim transcription of spontaneous formal speech from a period when such records are exceptionally rare.

During the 1830s, the well-known Hansard publication compiled debates from newspaper reports and generally paraphrased speeches. The Mirror operated differently (Barrow 1828, Jordan 1931, McBath 1970). It employed shorthand writers who transcribed speeches directly in the chamber, including the young Charles Dickens (Vice 2018), and who captured not only the words spoken but also paralinguistic features: audience reactions such as "(Laughter)", "(Hear, hear!)", and "(Cries of 'No, no!')", as well as speaker interruptions (Alexander 2023). As Jupp (1998: 204) says, the length of speeches in the Mirror match contemporaneous speech timings at a normal speaking pace, as an indicator of transcription detail. Prime Minister W.E. Gladstone noted in Parliament that for the early 1830s, the Mirror was "the primary record, and not Hansard's Debates, because of the greater fulness" [sic] it achieved (HC Deb 20 April 1877, c1576-77). The corpus analysed here has been created using a semi-automated pipeline involving AI-assisted OCR of original volumes, speaker identification based on contemporary title and naming patterns, extraction of session metadata, and preservation of paralinguistic markers. At present the corpus covers several volumes spanning the Reform Act debates, with speeches tagged for speaker, date, and session.

Preliminary findings show substantial differences between the Mirror and Hansard reports of identical debates (and, later, outright copying of the Mirror by Hansard). The corpus contains several hundred instances of marked paralinguistic features, enabling analysis of audience response patterns, interruption structures, and the performative dimension of parliamentary oratory. The corpus therefore addresses a gap in our sources for the history of English: we tend to have limited sources of extended spontaneous formal speech from earlier centuries preserving discourse features, with the notable exception of data such as the Old Bailey Corpus (Huber 2007, Archer 2014). The Mirror corpus therefore provides unusual access to turn-taking patterns, discourse markers in formal contexts, and speaker-audience interaction in the pre-recording era.

The paper presents some initial quantification of paralinguistic marker frequencies in the corpus, compares verbatim versus summary reporting through parallel examples of the Mirror corpus and the Hansard Corpus (Alexander and Davies 2015), and demonstrates potential applications for historical pragmatics and discourse analysis. As a work-in-progress, the paper discusses ongoing corpus expansion and annotation development, whilst es-

tablishing the value of this underutilised historical source for research into parliamentary discourse during a significant period of political reform.

References

Alexander, M. (2023). 'Speech in the British Hansard'. In Minna Korhonen, Haidee Kotze & Jukka Tyrkkö (eds.) Exploring Language and Society with Big Data: Parliamentary discourse across time and space. Amsterdam: John Benjamins. 17-53.

Alexander, M & Davies, M. (2015). The Hansard Corpus, 1803-2005. http://www.english-corpora.org/hansard

Archer, D. (2014), Historical Pragmatics: Evidence from the Old Bailey. Transactions of the Philological Society, 112: 259-277.

Barrow, JH. (1828). Prospectus. In John Henry Barrow (ed.) The Mirror of Parliament vol. 1. London: Winchester and Varnham.

Huber, M. (2007). The Old Bailey Proceedings, 1674-1834: Evaluating and annotating a corpus of 18th- and 19th-century spoken English. In Anneli Meurman-Solin & Arja Nurmi (eds.), Annotating variation and change, Helsinki: VARIENG. https://varieng.helsinki.fi/series/volumes/01/huber/

Jordan, HD. (1931). The Reports of Parliamentary Debates, 1803-1908. Economica 34. 437-49.

Jupp, P. (1998). British Politics on the Eve of Reform: The Duke of Wellington's Administration, 1828–30. London: Palgrave Macmillan.

McBath, JH. (1970). Parliamentary reporting in the nineteenth century. Communications Monographs, 37(1).

Vice, J. (2018). Charles Dickens and Gurney's Shorthand: 'That savage stenographic mystery'. Language and History 61. 77-93.

LLM vs. Human Language — WIP [2] E314 • 16:00–17:00

Towards a distinction between LLM-generated language and human language: Authenticity of corpus compilation in the age of artificial intelligence

WIP

Xiao Zhang (Xi'an International Studies University, China, People's Republic of)

Authenticity has been one of the foundational criteria for corpus compilation since the very beginning, requiring that a corpus represent genuine language use (Sinclair 1991, 1996; Hoffmann 2007; McEnery & Hardie 2012; McEnery & Brookes 2022). However, the application of large language model (LLM) could threaten this principle since 2022, particularly for corpora which represent web language, by flooding communicative spaces with AI-generated text. This compromises a corpus's capacity to function as a reliable record of human language. Current methods for LLM-generated text detection include black-box detection (e.g., Tang

et al. 2023), white-box detection (e.g., Verma et al. 2023), benchmarking method (e.g., He et al. 2023) and zero-shot method (e.g., Mao, et al. 2024). Some of the detection models embedded linguistics features, such as type-token ratio, lexical complexity, sentence structure complexity and stylistic elements (e.g., Lee et al. 2021; Guo et al. 2023). But these methods usually emphasized the distribution of features in the LLM-generated texts while overlooked a critical distinction: human language and AI-generated language exhibit systematically different types of linguistic "problems". To address this gap, this study proposes a detection method that incorporates analyses of lexical, syntactic, and textual anomalies.

The study hypothesize that LLM-generated texts can be distinguished from human language by three types of linguistic variables. The first variable is misspellings. Authentic texts often contain spellings errors because usually no correction would be adopted in corpus compilation, whereas LLM-generated texts are typically orthographically perfect. The second variable is syntactic completeness, which is proposed for the reason that human language often use fragments and elliptical structures while the LLM-generated sentences tend to be syntactically complete. The third one is keyness. Hypothetically, key words and key n-grams in LLM-generated texts may exhibit an abnormally high keyness scores since LLM models would overuse topic-related items in response to their prompts.

This study seeks to answer: 1) To what extent do LLM-generated texts fail to replicate these features of authentic human language? 2) How effective are these three variables in detecting LLM-generated texts?

To operationalise these variables, the English dataset of HC3 PLUS (95,000 training samples, 10,000 validation samples, and 38,000 test samples) is used as the LLM-generated texts dataset in this study. Authentic language dataset is drawn from a composite of the BNC 1994, the BNC 2014, and COCA. Misspellings are identified via #LancsBox with reference to the Birkbeck Spelling Error Corpus. Syntactic completeness was measured using confirmatory factor analysis. Keyness is analyzed for single words (#LancsBox) and n-grams (Extended Keyness Analysis Tool). A binary logistic regression tests the effects of these variables across the two datasets.

Preliminary results indicate that LLM-generated texts are characterized by fewer misspellings, more syntactically complete sentences, and more key items with high keyness scores than authentic human language. The findings incorporating such linguistic features that indicate anomalies can significantly improve detection methods, which in addition aspires to aide in the preservation of authenticity in future corpus compilation.

References

Ädel, A. (2020). Corpus compilation. In M. Paquot & S. Th. Gries (Eds.), A practical handbook of corpus linguistics (pp.3–24). Springer. Brezina, V. & Platt, W. (2025). #LancsBox X [Computer Software]. Lancaster University, http://lancsbox.lancs.ac.uk.

Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., & Wu, Y. (2023). How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection. arXiv. https://arxiv.org/abs/2301.07597

He, X. L, Shen, X. Y., Chen,Z. Y., Backes, M., & Zhang, Y. (2023). MGTBench: Benchmarking machine-

generated text detection. In Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security (pp. 2251-2265).

Hoffmann, S. (2007). From web page to mega-corpus: The CNN transcripts. In Hundt, M., Nesselhauf, N. & C. Biewer (Eds.), Corpus linguistics and the web (pp.69-85). Rodopi.

Lee, B. W., Yoo S. J., & Lee, J. H. (2021). Pushing on text readability assessment: A transformer meets handcrafted linguistic features. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (pp. 10669–10686), EMNLP 2021, Virtual Event.

Mao, C. Z., Vondrick, C., Wang, H. & Yang, J. F. (2024). Raidar: geneRative AI Detection via rewriting. arXiv. https://arxiv.org/abs/2401.12970

McEnery, T., & Hardie, A. (2012). Corpus linguistics: Method, theory and practice. Cambridge University Press.

McEnery, T., & Brookes, G. (2022). Building a written corpus: what are the basics? In O'Keeffe, A. & McCarthy, M. J. (Eds.), The Routledge handbook of corpus linguistics (pp. 35-47). Routledge.

Sinclair, J. (1991). Corpus, concordance, collocation. Oxford University Press.

Sinclair, J. (1996). Preliminary recommendations on corpus typology (Technical report. EAGLES (Expert Advisory Group on Language Engineering Standards). www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html .

Tang, R.X., Chuang, Y., & Hu, X. (2023). The science of detecting LLM-generated texts. arXiv. https://arxiv.org/abs/2303.07205

Verma, V., Fleisig, E., Tomlin, N., & Klein, D. (2023). Ghostbuster: Detecting text ghostwritten by large language models. CoRR, abs/2305.15047.

Zhang, X. (2025). Extended Keyness Analysis Tool (V1.0).

Corpora

British National Corpus (1994). http://www.natcorp.ox.ac.uk/

British National Corpus 2014. http://cass.lancs.ac.uk/bnc2014/

Birkbeck spelling error corpus (1980). https://ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/0643

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

Su, Z., Wu, X., Zhou, W., Ma, G. Y., & Hu, S. L. (2024). HC3 Plus: A semantic-invariant human ChatGPT comparison corpus. arXiv. https://arxiv.org/abs/2309.02731

Artificial registers and disciplinary voice: modelling human and AI-generated scientific writing

WIP

Daniele Polizzi, (University of Bologna, Italy)

Large language models (LLMs) are increasingly populating the field of scientific writing. Recent accounts document their influencing lexical choices across titles, abstracts and bodies of STEM research papers in particular (Zanotto & Aroyehun, 2024; Kobak et al., 2025;

Geng & Trotta, 2024), a trend that is more evident when these are authored by non-native speakers of English (Liu & Bu, 2024). While disciplinary registers historically reflect distinct epistemic and rhetorical norms (Biber & Conrad, 2019), and progressively diverge through specialization and diversification (Teich et al., 2015), an overreliance on AI-assisted text production may introduce new register ecologies that do not necessarily overlap with human tradition: a compressed space of language production that is statistically motivated and form-oriented rather than intent-driven (Bender & Koller, 2020) reducing variation beyond the scope of discrete words (Kwok et al., 2025). To address this issue, this work-in-progress study investigates whether LLM-generated academic prose preserves or erodes disciplinary distinctiveness through a register perspective, both within and across full-length texts.

The analysis focuses on three contiguous but stylistically distinct domains, namely general linguistics, computational linguistics, and computer science, representing a continuum from humanities to hard-science communication. This design is adopted to temper topical bias and allows register differences to emerge primarily from disciplinary communicative norms, ranging from interpretative through methodological to procedural discourse types. Within each discipline, twenty open-access, multi-authored, full-length research articles have been so far collected from journals with explicit AI-use disclosure policies following systematic random sampling. The selection included papers published from late 2022 onwards to minimize the risk of them being used as a basis to train foundation models. Each article is segmented into five functional sections mapped from heterogeneous labels to enable intra-article comparison: Background, Methods, Results, Discussion and Conclusion.

AI-generated counterparts are produced using two models differing in size (Qwen3-8B and Qwen3-32B) and two prompting regimes operationalizing different degrees of human–machine textual hybridity, following Terryn & de Lhoneux (2024) and Schepens et al. (2023): chain-of-thought, providing context through article metadata; and continuation, including section-initial sentences. Prompts are administered sequentially to ensure generation of comparable token counts.

Quantitative features selection follows Neumann & Evert (2021) and Teich et al. (2015) on the linguistic correlates of field, tenor and mode, and is performed at the section, document and discipline level. Variables cover lexico-grammatical (lexical density, noun–verb ratio, PoS distributions, modal verbs, frequent 3- and 4-gram bundles) syntactic (mean sentence length, subordination ratio, passives) and discourse-pragmatic (conjunction types, stance verbs) indicators. Multifactorial statistical analyses will test whether AI texts closely mirror human academic conventions (cf. Berber-Sardinha, 2024; Yun et al., 2023), or whether they exhibit lower intra- and inter-disciplinary dispersion, pointing to a hybrid artificial register marked by convergence and simplification.

References

Bender, E.-M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In Proceedings of the 58th annual meeting of the association for computational linguistics, 5185–5198.

Biber, D., & Conrad, S. (2019). Register, genre, and style. Cambridge University Press.

Geng, M., & Trotta, R. (2024). Is ChatGPT Transforming Academics' Writing Style?. arXiv. https://arxiv.org/abs/2404.08627

Kobak, D., González-Márquez, R., Horvát, E.-Á., & Lause, J. (2025). Delving into LLM-assisted writing in biomedical publications through excess vocabulary. Science Advances, 11(27), eadt3813.

Kwok, H.-L., Shi, Y., Xu, H., Li, D., & Liu, K. (2025). GenAI as a translation assistant? A corpus-based study on lexical and syntactic complexity of GPT-post-edited learner translation. System, 130, 103618.

Liu, J., & Bu, Y. (2024). Towards the relationship between AIGC in manuscript writing and author profiles: evidence from preprints in LLMs. arXiv. https://arxiv.org/abs/2404.15799

Neumann, S., & Evert, S. (2021). A register variation perspective on varieties of English. In Soane, E. & Biber, D. (Ed.), Corpus-based approaches to register variation (pp. 143–178). John Benjamins Publishing Company.

Sardinha, T.-B. (2024). AI-generated vs human-authored texts: A multidimensional comparison. Applied Corpus Linguistics, 4(1), 100083.

Schepens, J., Marx, N., & Gagl, B. (2023). Can we utilize large language models (LLMS) to generate useful linguistic corpora. A case study of the word frequency effect in young German readers. Open mind: discoveries in cognitive science, 9, 1597–1656.

Teich, E., Degaetano - Ortlieb, S., Fankhauser, P., Kermes, H., & Lapshinova - Koltunski, E. (2016). The linguistic construal of disciplinarity: A data - mining approach using register features. Journal of the Association for Information Science and Technology, 67(7), 1668–1678.

Terryn, A.-R., & de Lhoneux, M. (2024). Exploratory study on the impact of English bias of generative large language models in Dutch and French. In Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval)@ LREC-COLING 2024, 12–27.

Yun, H., Yi, E., & Song, S. (2023). Exploring AI-Generated English Relative Clauses in Comparison to Human Production. Journal of Cognitive Science, 24(4), 465–496.

Zanotto, S.-E., & Aroyehun, S. (2024). Human variability vs. machine consistency: A linguistic analysis of texts generated by humans and large language models. arXiv. https://arxiv.org/abs/2412.03025

Personal Anaphora in LLM-generated texts: Is AI language a new L2?

WIP

Xiao Zhang (Xi'an International Studies University, China, People's Republic of) & Shuang Liang (Guangdong Technion-Israel Institute of Technology, China, People's Republic of)

Personal anaphora refers to the use of a pronoun or pro-forms to refer back to a previously mentioned noun or noun phrase, the antecedent, in discourse. Its primary purpose in human language is to avoid repetition and make language flow more smoothly. Research on personal anaphora markers (e.g., he, she, it) in L2 English reveals systematic challenges beyond grammatical accuracy. Research on L2 English has consistently shown that even advanced learners produce non-native-like patterns in pronoun selection and consistency, influenced by L1 transfer, cognitive load, or discourse-pragmatic differences. A key area of

investigation is the overt subject pronoun requirement in English, contrasting with pro-drop languages like Spanish or Italian. Learners from these L1 backgrounds frequently overuse subject pronouns, leading to redundancy in contexts where native speakers would use null or zero anaphora. Furthermore, learner-corpus-based studies also highlight issues with anaphoric ambiguity, where learners fail to maintain clear referential chains, and anaphoric underuse, where pronouns are omitted where required. These patterns underscore that mastering anaphora involves complex interactions between syntax, discourse, and pragmatics.

Human learners usually acquire the rules governing personal anaphora in L2 through instruction, interaction, and implicit learning. In contrast, large language models (LLMs) learn statistical patterns of association between pronouns and their likely antecedents from massive text data during training. This approach could be understood as a pattern-matching process in which LLMs learn surface-level co-occurrence relations between pronouns and antecedents, as well as distributional regularities in syntax, semantics, and discourse. LLMs and L2 learners may display observable similarities in using anaphora.

The present study proceeds from the hypothesis that LLMs' anaphora behavior more closely resembles L2 learner language than native speaker language. Specifically, LLMs, like L2 learners, may exhibit incorrect subject antecedent, ambiguous reference, or reduced sensitivity to discourse-pragmatic constraints on anaphora. Both in NLP and SLA, anaphora, a linguistic device relying on both syntactic knowledge and pragmatic inference, provides disambiguating information through establishing co-referential links. Against this background, the present study aims to determine whether the deployment of personal anaphora can serve as a defining indicator for identifying parallels between texts generated by LLMs and those produced by L2 learners of English.

To address this aim, instances of personal anaphora markers are retrieved from three complementary corpora: 1) The Uppsala Student English Corpus, a learner corpus comprising 1,489 essays, totalling 1,221,265 words; 2) The Louvain Corpus of Native English Essays (LOCNESS), a native English writing corpus, totalling 324,304 words; 3) the English dataset of Human ChatGPT Comparison Corpus (HC3), including 26,903 samples of ChatGPT-generated texts. Anaphora Resolution of this study is conducted by using AllenNLP.

The findings are anticipated to elucidate the extent of similarity in the deployment of personal anaphora across the three corpora: LLM-generated English texts vs. L2 English texts, and LLM-generated English texts vs. native English texts. Ultimately, this study investigates the proposition that AI-generated English may share greater linguistic similarities with learner English than with native speaker English, particularly in the domain of discourse-level referential coherence.

References

Bärenfänger, M., Goecke, D., Hilbert, M., Lüngen, H., & Stührenberg, M. (2008). Anaphora as an indicator of elaboration: A corpus study. Journal for Language Technology and Computational Linguistics, 23(2), 49–72.

Fox, B. A. (1987). Discourse structure and anaphora. Cambridge University Press.

Granger, S. (1998). The computer learner corpus: A versatile new source of data for SLA research. In Granger, S. (Ed.), Learner English on computer (pp. 3-18). Addison Wesley Longman.

Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., & Wu, Y. (2023). How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection. arXiv. https://arxiv.org/abs/2301.07597

Gundel J.K. & Abbott, B. (Eds.) (2019). The Oxford handbook of reference. Oxford University Press.

Joshi, M., Chen, D., Liu, Y., Weld, DS., Zettlemoyer, L., & Levy, O. (2020). SpanBERT: improving pre-training by representing and predicting spans. Transactions of the Association for Computational Linguistics, 64–77.

Kolhatkar, V., Roussel, A., Dipper, S., & Zinsmeister, H.(2018). Anaphora with non-nominal antecedents in computational linguistics: A Survey. Computational Linguistics, 44 (3), 547–612.

Lee, C., Jung, S., & Park, CE. (2017). Anaphora resolution with pointer networks. Pattern Recognition Letters, 95, 1-7.

Liu, R. & Nicol, J. (2010). Online Processing of Anaphora by Advanced English Learners. In Matthew T. Prior et al. (Eds.), Selected Proceedings of the 2008 Second Language Research Forum (pp. 150-165). Somerville, MA: Cascadilla Proceedings Project.

Mitkov, R. (2002). Anaphora resolution. Longman.

Poesio M. & Artstein, R. (2008). Anaphoric annotation in the ARRAU corpus. Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008) (pp. 1170–74). Paris: European Language Resource Association.

Poesio, M., Grishina, Y., olhatkar, V., Moosavi, N., Roesiger, I., Roussel, A., Simonjetz, F., Uma, A., Uryupina, O., Yu, J. T., & Zinsmeister, H. (2018). Anaphora resolution with the ARRAU corpus. In Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference (pp. 11–22), New Orleans, Louisiana. Association for Computational Linguistics.

Poesio, M., Yu, J. T., Paun,S., Aloraini, A., Lu, P.C., Haber, J., & Cokal, D. (2023). Computational models of anaphora. Annual Review of Linguistics, 9, 561-587.

Sileo, R.B., Cilibrasi, L., Heine, J., & Tsimpli, I.M. (2024). The role of aspect on anaphora resolution in English as a first and second language. Journal of the European Second Language Association, 8(1), 48–65.

Sorace, A., & Filiaci, F. (2006). Anaphora resolution in near-native speakers of Italian. Second Language Research, 22(3), 339–368.

Swanson, K. & Dekydtspotter, L. (2022). A Full Parse or a Shallow Structure in L2 An ERP Study of Anaphora in Successive-Cyclic Wh-movement in L1-Mandarin/L2-English. In Gong, Y. & Kpogo, F. (Eds.), Proceedings of the 46th annual Boston University Conference on Language Development (pp. 768-782). Cascadilla Press.

Wan, J., & Ren, H. (2024). The features, classification, and functions of event anaphora in Chinese news discourse. Advances in Education, Humanities and Social Science Research, 9, 81-85.

White, L. (2003). Second language acquisition and universal grammar. Cambridge University Press.

Xia, P. & Durme, BV. (2021). Moving on from OntoNotes: Coreference resolution model transfer. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 5241–56). Stroudsburg, PA: Association for Computational Linguistics.

Corpora

Human ChatGPT Comparison Corpus (HC3): https://github.com/Hello-SimpleAI/chatgpt-comparison-detection

The Uppsala Student English Corpus (USE): https://ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/2457

The Louvain Corpus of Native English Essays (LOCNESS): https://www.learnercorpusassociation.org/resources/tools/locness-corpus/

RHINE

MOSELLE

Friday
29 May 2026

Friday 29 May 2026

German Learner English E113 • 10:30–12:30

Vowel effects on rhoticity in German learners of English

FULL PAPER

Alina Waitzmann (University of Bamberg, Germany)

Rhoticity is a key feature distinguishing different varieties of English. In rhotic accents, such as General American (GA), the phoneme /r/ is realised in all phonetic contexts, whereas in non-rhotic accents, such as Standard Southern British English (SSB), /r/ is typically restricted to prevocalic positions. Yet, language-internal as well as language-external factors can influence the extent of rhoticity even within native speaker varieties. Research by Irwin & Nagy (2007) and Piercy (2012), for example, has shown that among others preceding vowels have an impact on rhoticity in Boston and Dorset English with the NURSE vowel eliciting the highest degree of rhoticity. The influence of preceding vowels has also been explicitly analysed twice in EFL varieties (Kang 2013, Li & Kabak 2017). Li and Kabak (2017) identify the preceding vowel as the most significant phonological factor for non-rhotic or rhotic realisations in Chinese speakers while Kang (2013) finds it insignificant for Korean speakers of English. This suggests that, besides the preceding vowel, the analysis needs to factor in the speakers' L1.

German can be considered variably rhotic in non-prevocalic positions with a tendency for r-vocalisation. In German, the preceding vowel's length, quality, and position are connected to the realisation of /r/ (Krech et al. 2009, Kautzsch 2017). Based on realisation patterns of non-prevocalic /r/ in German and previous research on rhoticity in native speakers of English, we hypothesise that the preceding vowel has an influence on the distribution of /r/ in the English of German learners. We further predict that /Vr/ segments will be most rhotic with the NURSE vowel and least rhotic with the LETTER vowel. Speech data were collected from 40 German undergraduate students of English recording a counterbalanced word list with /Vr/ segments containing the NORTH, START, NEAR, SQUARE, CURE, NURSE, and LETTER vowels. For each participant, the non-prevocalic /r/ tokens were analysed auditorily and acoustically. Acoustically, the presence of [ɹ] was evaluated based on F3–F2 with tokens classified as rhotic if the difference was below 950 Hz. Only tokens for which auditory and acoustic classifications agreed were included in the statistical analysis.

In order to contextualise the findings, the distribution of non-prevocalic /r/ was also assessed for every possible /Vr/ combination in German. Additionally, the speaker's general tendency towards rhotic or non-rhotic productions was controlled for by determining their apparent

English target variety. Findings suggest that rhoticity in German learners of English is conditioned by vowel context. Specifically, rhotic realizations are most frequent following the NURSE vowel and least frequent following the LETTER vowel, in line with initial predictions.

References

Irwin, P. & Naomi N. (2007). Bostonians /r/ speaking: A quantitative look at (R) in Boston. U. Penn Working Papers in Linguistics, 13(2), 135-147.

Kang, H.-S. (2013). Internal and external constraints on rhoticity in Korean English. The Sociolinguistic Journal of Korea, 21(2), 1-28.

Kautzsch, A. (2017). The attainment of an English accent: British and American features in advanced German learners. Inquiries in Language Learning 20. Frankfurt a. M.: Peter Lang Edition.

Krech, E.-M., Stock, E., Hirschfeld, U. & Anders, L.U. (eds.). (2009). Deutsches Aussprachewörterbuch. Berlin/New York: Mouton de Gruyter.

Li, Z. & Kabak, B. (2017). Rhoticity in Chinese English: An experimental investigation on the realization of the variant (r) in an Expanding Circle Variety. Alicante Journal of English Studies, 30, 61-91.

Piercy, C. (2012). A transatlantic cross-dialectal comparison of non-prevocalic /r/. University of Pennsylvania Working Papers in Linguistics, 18(2), 77-86.

/ˈdʒɜːmən/ or /ˈtʃɜːmən/: Sibilant devoicing in Southern German speakers of English

FULL PAPER

Alina Waitzmann & Julia Schlüter (University of Bamberg, Germany)

In this paper, we address one typical pronunciation challenge faced by learners of English from Southern Germany, which is however rarely mentioned in studies of learner phonology (e.g. Gut 2009, Kautzsch 2017, Sönning 2020) or pronunciation guides (e.g., Eckert & Barry 2005, Schmitt 2016): the voiceless realisation of voiced sibilants. While the voiced sibilant /z/ is present and systematically available in Standard German, it is systematically absent in southern dialects (Zehetner 1982, Burgschmidt & Götz 1972). Another systematic gap in southern German varieties concerns the voiced sibilants /ʒ/ and /dʒ/, which are marginally present in the Standard German phoneme inventory, occurring only in loanwords (Żygis et al. 2012). Among speakers of German, proficiency in this subsystem of the standard phonology varies widely, as lack of voicing is hardly overtly stigmatized.
We hypothesize that the absence of a systematic voicing contrast for sibilants in Southern German dialects results in negative transfer, with learners of L2 English producing the voiceless counterparts /s/, /ʃ/, and /tʃ/ in contexts where voiced sibilants would be required. Thus, we pose the following research questions:
(1) Which language-external factors influence the presence or absence of voicing?
(2) Can we discern system-internal implicational relationships, such that
a) speakers realising the voiced sibilants in German also realise them in English, but not

vice versa,
b) in English, speakers successfully realising the voiced sibilants word-finally also possess them word-medially, but not vice versa,
c) in English, speakers are more likely to retain voicing in one (or two) of the three sibilants than the other(s)?
The data stem from a corpus comprised of 250 recordings by German university students of English reading a specific diagnostic text. For a subset of speakers, L1 productions of (Standard German) /z/, /ʒ/, and /dʒ/ (dialectally realized as /s/, /ʃ/, and /tʃ/) are available, enabling us to assess the presence of the voicing feature in their L1 phoneme inventories. Tokens are analysed auditorily by two phonetically trained raters, with inter-rater reliability calculated using Cohen's kappa. In addition, an acoustic analysis is conducted, including the measurements of duration of voicing and duration of the entire phoneme. It is generally presumed that voiceless sibilants exhibit longer durations than their voiced counterparts (Fuchs et al. 2009, Żygis et al. 2012).
Based on previous research, we expect the following language-external factors to be relevant: the speakers' regional origin (Zehetner 1982), their gender (Weirich & Simpson 2015, Funk et al. 2025), and their time spent in an English-speaking country (Kautzsch 2017). We also anticipate a significant impact of final devoicing – a systematic process affecting all obstruents in German regardless of L1 variety (Sönning 2020). Preliminary (auditory) results suggest that the three sibilant types are strongly correlated, indicating that speakers tend to either master voiced sibilants as a subsystem of the English phoneme inventory, or to lack it. Finally, for the subset of speakers for whom recordings in both English and German are available, the effect of the presence or absence of voiced sibilants in their L1 seems to be confirmed.

References

Burgschmidt, E. & Götz, D. (1972). Kontrastive Phonologie Deutsch-Englisch und Mundartinterferenz. In A. James & B. Kettemann (eds.), Dialect Phonology and Foreign Language Acquisition, 32-47. Tübingen: Narr.

Eckert, H. & Barry, W. J. (2005). The phonetics and phonology of English pronunciation: A coursebook with CD-ROM. 2nd ed. Trier: Wissenschaftlicher Verlag.

Fuchs, S., Brunner, J. & Busler, A. (2009). Temporal and spatial aspects concerning the realizations of the voicing contrast in German alveolar and postalveolar fricatives. International Journal of Speech-Language Pathology, 9(1), 34-44. https://doi.org/10.1080/14417040601094315

Funk, R., Weirich, M., Simpson A. P. (2025). How sibilant spectra shape gender perception in prepubertal children: A voice morphing study. Interspeech 2025, 963-967. 10.21437/Interspeech.2025-125

Gut, U. (2009). Non-native speech: A corpus-based analysis of phonological and phonetic properties of L2 English and German. Frankfurt am Main: Lang.

Kautzsch, A. (2017). The attainment of an English accent: British and American features in advanced German learners. Berlin: Lang.

Schmitt, H. (2016). Teaching English pronunciation: A textbook for the German-speaking coun-

tries. Heidelberg: Winter.

Simpson, A. P. & Weirich, M. (2015). Gender-specific differences in sibilant contrast realizations in English and German, ICPhS 2015.

Sönning, L. (2020). Phonological variation in German learner English. University of Bamberg dissertation. https://doi.org/10.20378/irb-49135

Zehetner, L. (1982). 'Baierisches Englisch'. Muttersprachiger Dialekttransfer im Fremdsprachenerwerb am Beispiel regionalspezifischer Schwierigkeiten und Möglichkeiten für den Englischunterricht im bairischen Dialektraum. In A. James & B. Kettemann (eds.), Dialect Phonology and Foreign Language Acquisition, 32-47. Tübingen: Narr.

Żygis, M., Fuchs, S. & König L. L. (2012). Phonetic explanations for the infrequency of voiced sibilant affricates across languages. Laboratory Phonology, 3(2), 299-336. https://doi.org/10.1515/lp-2012-0016

Fluency across modes: The case of German learners of English

FULL PAPER

Bethany Stoddard (University of Bonn, Germany), Lisa-Christine Altendorf (University of Bonn, Germany) & Lea Bracke (University of Bamberg, Germany)

The triad of complexity, accuracy, and fluency (CAF) has gained much traction in Learner Corpus Research for mapping the multi-faceted nature of second language proficiency. This study focuses on individual patterns of fluency in spoken and written L2 English production among German secondary school learners, with a particular focus on the role of anxiety. While previous studies have examined accuracy and complexity across modalities (i.e., Foster & Wigglesworth, 2016; Biber et al., 2024), few studies have explored fluency in cross-modal comparisons (e.g. Bilge et al., 2022; Suzuki & Révész, 2023), primarily due to the rarity of parallel spoken and written data from the same learners. Anxiety has typically been shown to negatively affect spoken (e.g. Pérez Castillejo, 2019; Bielak, 2025) and written (e.g. Güvendir & Uzen, 2023; Zabihi et al., 2020) CAF measures, but more rarely, it may also have a facilitative effect. Macayan et al. (2018) found a facilitative effect of anxiety on L2 writing performance, but a negative effect on speaking. Thus, examining both modes allows us to better understand the role of anxiety on each mode of production.

We posit that learners will exhibit similar fluency levels across modalities: those demonstrating low fluency in writing will also perform at the lower end in speaking, and vice versa. This cross-modal consistency would suggest that fluency reflects underlying cognitive processes that transcend mode-specific production constraints (see Housen, Kuiken & Vedder, 2012; De Jong et al., 2015). Furthermore, we hypothesize that exam anxiety will negatively affect fluency in speech more than in writing.

The study utilizes written and spoken data from the YGLE Corpus (Bracke et al., 2024), which includes cross-sectional data from young English learners in German secondary schools, as well as an extensive set of metadata. As the corpus is still being compiled, we use a subset

of data from around 300 learners across multiple grades (7, 9, and 11) and school types (Realschule and Gymnasium). Spoken fluency is operationalized via speech rate (number of syllables/speaking time including pauses), articulation rate (number of syllables/speaking time excluding pauses), and filled pause ratio. Written fluency is operationalized through product-oriented measures adapted from oral fluency research: words per T-unit (W/T), and words per clause (W/C), alongside total words per text (see Bilge et al., 2022; Barrot & Gabinete, 2019). Test anxiety is measured via the Fragebogen zur Leistungsmotivation (Achievement Motivation Questionnaire; Petermann & Winkel, 2015).
Fluency measures will be converted to z-scores to ensure comparability between the written and spoken measures. Using linear mixed-effects modeling, we will examine the influence of test anxiety on fluency in each mode and the relationship between individuals' spoken and written fluency while controlling for individual differences (IQ, grade level, school type, and region). By examining fluency across modalities, we aim to determine whether fluency represents a generalized linguistic skill or remains fundamentally mode-dependent, with implications for pedagogy and assessment in L2 contexts.

References

Biber, D., Larsson, T. & Hancock, G. (2024). Dimensions of Text Complexity in the Spoken and Written Modes: A Comparison of Theory-Based Models. Journal of English Linguistics, 52(1), 65-94. https://doi.org/10.1177/00754242231222296 .

Bielak, J. (2025). To what extent are foreign language anxiety and foreign language enjoyment related to L2 fluency? An investigation of task-specific emotions and breakdown and speed fluency in an oral task. Language Teaching Research, 29(3), 911-941.

Bilge, H., & Kalenderoğlu, İ. (2022). The relationship between reading fluency, writing fluency, speaking fluency, reading comprehension, and vocabulary. Education and Science, 47(209), 25-53.

Bracke, L., Stoddard, B., Fuchs, R., Rosen, A. & Werner, V. (2024, March 21–22). Introducing the Corpus of Young German Learner English [Conference presentation]. Norddeutsches Linguistisches Kolloquium, Hannover, Germany.

de Jong, N. H., Pacilly, J. & Heeren, W. (2021). PRAAT scripts to measure speed fluency and breakdown fluency in speech automatically. Assessment in Education: Principles, Policy & Practice, 28(4), 456–476. https://doi.org/10.1080/0969594X.2021.1951162

Foster, P. & Wigglesworth, G. (2016). Capturing Accuracy in Second Language Performance: The Case for a Weighted Clause Ratio. Annual Review of Applied Linguistics, 36, 98–116. doi: 10.1017/S0267190515000082

Güvendir, E. & Uzun, K. (2023). L2 writing anxiety, working memory, and task complexity in L2 written performance, Journal of Second Language Writing, 60, Article 101016, https://doi.org/10.1016/j.jslw.2023.101016 .

Petermann, F. & Winkel, S. (2015). Fragebogen zur Leistungsmotivation für Schüler der 7. bis 13. Klasse. Pearson Harcourt.

Pérez Castillejo, S. (2019). The role of foreign language anxiety on L2 utterance fluency during a final exam. Language Testing, 36(3), 327-345. https://doi.org/10.1177/0265532218777783 .

Suzuki, S. & Révész, A. (2023). Measuring speaking and writing fluency: A methodological synthe-

sis focusing on automaticity. In Practice and automatization in second language research (pp. 235-264). Routledge.

Zabihi, R., Mousavi, S. H. & Salehian, A. (2020). The differential role of domain-specific anxiety in learners' narrative and argumentative L2 written task performances. Current Psychology, 39(4), 1438-1444. https://doi.org/10.1007/s12144-018-9850-6 .

Mapping linguistic complexity in German learners' written English productions

FULL PAPER

Lisa-Christine Altendorf (University of Bonn, Germany)

Linguistic complexity is a key element for capturing differences in language proficiency and tracking its development (Ortega, 2012). Complexity has mostly been assessed using syntactical and lexical complexity, with dimensions such as morphological complexity (De Clercq & Housen, 2016), phraseological complexity (Pacquot, 2017; Vandeweerd, Housen & Pacquot, 2023) and grammatical complexity (Biber et al., 2022) recently gaining traction.
Multidimensional models of linguistic complexity emphasize that complexity is not a monolithic construct (Bulté & Housen, 2012; Norris & Ortega, 2009). Rather, it manifests across these different interconnected domains, each contributing to the overall sophistication and quality of learner production. Biber's (1992, 2014) multidimensional analyses illustrate how grammatical and phrasal patterns co-occur in context-dependent ways, while research in L2 writing (Lu, 2011; Kyle et al., 2016; Vyatkina, 2012) has shown that lexical sophistication, nominal elaboration, and subordination develop along partly independent trajectories. This theoretical grounding provides a rationale for exploring not only individual indices of complexity but also their interrelations, offering a more complete account of how language learners construct increasingly complex written language over time.
The central research question guiding this study thus is: How do the different dimensions of linguistic complexity interact at different proficiency levels in learner writing? It is hypothesized that while overall growth will be observed across all dimensions, there will be trade-offs between the dimensions of complexity at different proficiency levels, reflecting the cognitive constraints and developmental trajectories of language learners.
Data are drawn from the YGLE Corpus (Bracke et al., 2024), comprising writing samples from approximately 300 English learners in German secondary schools across grades 7, 9, and 11 in two different tracks. Lexical complexity is measured using the Tool for the Automatic Analysis of Lexical Sophistication (TAALES; Kyle, Crossley, & Berger, 2018). Grammatical and syntactic complexity are analyzed with Kyle's (2025) Lexicogrammatical Tagger based on Biber's (1992, 2014) multidimensional model, while phraseological complexity is operationalized through mean pointwise mutual information (MI) scores of adjective–noun and verb–object combinations (Paquot, 2017). Morphological complexity is measured using the Type/Family ratio (Horst & Collins, 2006), Inflectional Diversity (Malvern et al., 2004), and the Morphological Complexity Index (Brezina & Pallotti, 2016).

Principal Component Analysis (PCA) is applied within each grade group to uncover how different measures cluster and vary with proficiency. While PCA is appropriate for mapping the latent structure among interrelated complexity dimensions, a multiple regression modeling will further test relationships and developmental predictors.
By examining multiple complexity dimensions simultaneously, this corpus-based study contributes to our understanding of how linguistic complexity develops in instructed second language writing. The findings will have implications for language assessment, curriculum design, and our theoretical understanding of the relationship between different aspects of linguistic complexity in second language development.

References

Biber, D. (1992). On the Complexity of Discourse Complexity: A Multidimensional Analysis. Discourse Processes, 15(2), 133-163. https://doi.org/10.1080/01638539209544806

Biber D. (2014). Using multidimensional analysis to explore cross-linguistic universals of register variation. Languages in Contrast, 14(1), 7–34.

Biber, D., Gray, B., Staples, S., & Egbert, J. (2022). The register-functional approach to grammatical complexity: Theoretical foundation, descriptive research findings, application. Routledge. https://doi.org/10.4324/9781003087991

Bracke, L., Stoddard, B., Fuchs, R., Rosen, A., & Werner, V. (2024, March 21–22). Introducing the Corpus of Young German Learner English [Conference presentation]. Norddeutsches Linguistisches Kolloquium, Hannover, Germany.

Brezina, V., & Pallotti, G. (2016). Morphological complexity in written L2 texts. Second Language Research, 35(1), 99-119. https://doi.org/10.1177/0267658316643125 .

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken & I. Vedder (Eds.), Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA (pp. 21 – 46). John Benjamins.

De Clercq, B., & Housen, A. (2016). The development of morphological complexity: A cross-linguistic study of L2 French and English. Second Language Research, 35(1), 71-97. https://doi.org/10.1177/0267658316674506

Ellis, N., Römer, U., & O'Donnell, M. (2013). Usage-based approaches to language acquisition and processing: Cognitive and corpus investigations of construction grammar. Mouton de Gruyter.

Horst, M., & Collins, L. (2006). From Faible to Strong: How Does Their Vocabulary Grow? Canadian Modern Language Review, 63(1), 83–106.

Housen, A., Kuiken, F., & Vedder, I. (2012). Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA. John Benjamins.

Kyle, K. (2025). Lexicogrammatical Tagger (LxGrTgr). https://github.com/kristopherkyle/LxGrTgr

Kyle, K., Crossley, S. A., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. Behavior Research Methods, 50(3), 1030–1046.

Lu, X. (2011). A corpus-based analysis of syntactic complexity in L2 writing. International Journal of Corpus Linguistics, 15(4), 474–496.

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). Lexical diversity and language develop-

ment: Quantification and assessment. Palgrave Macmillan.

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

Ortega, L. (2012). Interlanguage complexity: A construct in search of theoretical renewal. Language Learning Supplement, 62(1), 1–36.

Paquot, M. (2017). The phraseological dimension in interlanguage complexity research. Second Language Research, 35(1), 121–145.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. Applied Linguistics, 30(4), 510–532.

Tomasello, M. (2003). Constructing a language: A usage-based theory of language acquisition. Harvard University Press.

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(4), 787–811.

Vyatkina, N. (2012). The development of second language writing complexity in groups and individuals: A longitudinal learner corpus study. Modern Language Journal, 96(4), 572–594. http://hdl.handle.net/1808/26403

Cottontails, ladyfingers, and laptops: What are the roles of metaphor and metonymy in English exocentric compounds?

FULL PAPER

Qingnan Meng (Dalian Maritime University, China, People's Republic of), Weihua Luo (Dalian Maritime University, China, People's Republic of) & Martin Hilpert (University of Neuchatel, Switzerland)

Exocentric compounds such as cottontail or laptop carry non-compositional meanings that cannot be deduced from the meanings of their parts. It has been argued that exocentric compounds can be classified as metaphorically/metonymically endocentric (Plag 2018). This paper aims to work out the roles of metaphor and metonymy in exocentric compounds on the basis of corpus data and distributional semantic methods, using the COCA and enTenTen21. Specifically, we aim to test whether Plag's (2018) semantic right-headedness hypothesis hold for English exocentric compounds. We operationalize headedness through distributional semantic methods, creating semantic vector spaces for 60 exocentric compounds and their components (see Table 1 in the attached pdf file).
Our results show that only a few compounds in the dataset (e.g. cottontail, ladyfinger, humpback, laptop) clearly exhibit exocentric distributional features (see Figure 1 in the attached pdf file). For these compounds, the two source words show a semantic behavior that is very different from that of the compound. For most compounds in the dataset,

however, their context vector tokens overlap to a large extent with those of at least one source word, suggesting their endocentric nature. Regarding headedness, our results indicate that endocentricity does not necessarily imply right-headedness. Based on the distributional features of the MDS plots, left-headed, right-headed and double-headed exocentric compounds appear equally plausible in our dataset.
In terms of the relationship between metaphor-related compounds (e.g. ladyfinger representing a pastry with a finger-like shape) and metonymy-related compounds (e.g. paleface denoting a person with a pale face), the average classification accuracy of the multinomial logistic regression model, using MDS coordinates as two predictors, suggests that they lie on a continuum from endocentricity to exocentricity. Metaphor-related compounds tend to lean towards the exocentric end, while metonymy-related compounds are closer to the endocentric end (see Figure 2 in the attached pdf file). This contrasts with Halupka-Rešetar & Lalić-Krstin's (2009) proposal on the continuum of Serbian blends, where metaphor-related blends are closer to the endocentric end, likely due to cross-linguistic or methodological differences between blending and compounding processes.
Overall, our findings offer new empirical insights into the semantic structure of English compounds that help to advance cognitive semantic theory. In addition, it may also shed light on further verification of the relative positions of metaphor- and metonymy-related compounds on the continuum from a typological perspective.

References

Meng, Q. & Hilpert, M. (2024). Measuring the semantic headedness of English blends with token-based semantic vector space modeling: A corpus-based study. Digital Scholarship in the Humanities, 39(4), 1075-1091. https://doi.org/10.1093/llc/fqae050

Evaluation of different approaches to metaphor identification in large corpora: A case study on spoken interactions involving pre-teens in England

FULL PAPER

Alice Deignan (University of Leeds, United Kingdom), Elena Semino (Lancaster University, United Kingdom), Sarah Daniel (University of Leeds, United Kingdom), Eleanor Field (University of Birmingham, United Kingdom) & Jeannette Littlemore (University of Birmingham, United Kingdom)

It is well established that metaphor is a central device for communication and thinking. Large corpora of naturally-occurring linguistic data are often analysed both for the purposes of theory development and to investigate the role of metaphors in framing particular topics and experiences in different discourse contexts. Such analyses, however, have to address the challenge of how to identify linguistic metaphors in datasets that are too large to be analysed manually through one of several established identification procedures such as MIP (Pragglejaz, 2007) or MIPVU (Steen et al, 2010).
There have been various approaches to this challenge. Charteris-Black's (2004) small-to-

large corpus technique involves manually annotating a sample then searching the full corpus for the metaphors identified. Cameron and Deignan (2003) searched a small corpus for linguistic devices that often function to introduce metaphors, such as kind of, really and imagine, and concordanced these in a larger corpus to find metaphors. Deignan and Semino (2010) used the software package WMatrix to identify the most frequent semantic fields in corpora, and where these seemed anomalous with the central topics, they tended to find metaphorical uses. Berber-Sardinha (e.g. 2012) has used corpora that have been manually marked for metaphor to train a program to predict the probability of each word in a larger corpus being metaphorically used. Key Words and Wordlists have also been used as starting points (Berber-Sardinha, 2012; Deignan et al 2019). More recently, generative AI tools have been tested for their performance in metaphor identification. In particular, Fuoli et al. (2025) have explored different approaches to the use of Large Language Models to identify metaphor, with some success. They additionally found that discrepancies between human and LLM coding tended to occur in 'grey' areas already known to metaphor scholars. Overall, however, previous work has tended to focus on written language produced by adults.

This paper compares the results of different approaches to metaphor identification by examining a particularly challenging and under-studied type of data: informal spoken language produced by children aged 9-13 years, in the context of group interviews conducted in around 12 schools in England. Specifically, we compare the findings of three approaches to metaphor identification: manual analysis of a random sample from the corpus, involving consensus through reliability testing among the co-authors of this paper; manual analysis of concordances of the most frequent lexical items and semantic domains; and a fine-tuning AI method inspired by Fuoli et al. (2025) but involving a secure (and hence ethically acceptable) LLM. We discuss the advantages and disadvantages of each approach, in terms of reliability, coverage and efficiency. We show how LLMs can be exploited ethically alongside both manual analysis and more transparent if more time-consuming corpus-based techniques. Finally, we build on Fuoli et al. (2025) by considering the implications of discrepancies in human and AI coding in the specific case of informal spoken interactions involving 9 to 13 year old children.

References

Berber Sardinha, T. 2012. An assessment of metaphor retrieval methods. In MacArthur, F., Oncins-Martínez, J., Sánchez-Gárcia, M., & Piquer-Píriz, A. Metaphor in use: Context culture and communication, John Benjamins.

Cameron, L. & Deignan, A. 2003.Combining large and small corpora to investigate tuning devices around metaphor in spoken discourse. Metaphor and Symbol, 18(3), 149-160.

Charteris-Black, J. 2004. Corpus approaches to critical metaphor analysis. Palgrave.

Deignan, A. & Semino, E. 2010. Corpus techniques for metaphor analysis. In Cameron, L., Maslen, R. (eds) Metaphor analysis: research practice in applied linguistics, social sciences, and the humanities. Equinox.

Deignan, A., Semino, E. & Paul, S. 2019.Metaphors of climate science in three genres: Research

articles, educational texts and secondary school student talk. Applied Linguistics, 40(2), 379-403.

Fuoli et al (2025, under review). Metaphor identification using large language models: A comparison of RAG, prompt engineering and fine tuning.

Pragglejaz group, 2007. MIP: A method for identifying metaphorically-used words in discourse. Metaphor and Symbol. 22(1), 1-39. Steen, G., Dorst, A., Herrmann, J., Kaal, A., Krennmayr, T. & Pasma, T. 2010. A method for linguistic metaphor identification. John Benjamins.

AI-Assisted Corpus Analysis of Spatio-temporal Metaphors

WIP

Águeda Salmerón Hortelano (Universidad de Murcia, Spain)

The conceptualization of TIME is frequently framed through its metaphorical projection onto the concrete domain of SPACE (Lakoff & Johnson, 1980). Within this mapping, two major construals are typically distinguished: the Ego-moving (EM) and the Time-moving (TM) metaphors (see Casasanto & Boroditsky, 2008; Gentner et al., 2002; Illán Castillo, 2024; Loerman & Milfont, 2018; Valenzuela Manzanares & Illán Castillo, 2022). In the EM construal, the experiencer functions as a dynamic entity progressing through temporal space (e.g., "We are heading into winter"), whereas the TM pattern depicts time itself as the entity in motion, advancing towards a static observer (e.g., "Winter is approaching").

Although the emotional and cognitive implications of these metaphors have received attention in psycholinguistic research (Hauser et al., 2009; Richmond et al., 2012; Ruscher, 2011), their spontaneous use in natural language remains insufficiently examined. Prior corpus-based investigations (Feist & Duffy, 2020; McGlone & Pfiester, 2009) have offered preliminary observations but continue to face methodological constraints. This research addresses two guiding questions: (i) what semantic and affective patterns emerge from EM and TM metaphors from large-scale corpora? and (ii) how can computational methods account for their asymmetry in frequency and structure?

To address these aims, a dual-corpus approach has been adopted. The enTenTen21 corpus from SketchEngine (Kilgarriff et al., 2014) provides a broad reference, while the NewsScape v5 dataset from CQPweb (Hardie, 2014) supports the construction of an adhoc subcorpus that reflects natural metaphor use in broadcast discourse. The analysis follows the corpus query of the motion verb "approach", repeatedly validated in previous literature for its cross-metaphor salience, and focuses on occurrences where temporal entities function either as the subject (TM) or as the object (EM) (see Margolies & Crawford, 2008; Soriano & Piata, 2022).

Building upon this corpus foundation, a semi-automated processing framework was developed through custom Python scripts co-created with ChatGPT (OpenAI, 2025). These scripts employ NLP libraries such as SpaCy, NLTK and pandas to streamline several analytical stages: (i) quantified extraction of verbal tense per metaphor type, revealing that EM metaphors overwhelmingly favour the present simple (e.g., "As we approach summer"), whereas TM metaphors display a more even distribution across tenses, with the present continuous and

gerund forms being particularly salient (e.g., "With the anniversary approaching"); (ii) extraction of collocational patterns surrounding time entities (e.g., identifying "his 40th birthday" from "birthday" queries); or (iii) semi-automatic tagging and trend analysis using ontology methods derived from the Time Event Ontology (Cox & Little, 2022) to explore temporal conceptualisation. Crucially, it illustrates how AI-assisted scripting enables researchers to remain actively in control of the analytical process while guiding the system to address theoretical inconsistencies inherent in abstract domains such as TIME.

Overall, this study contributes a methodological framework that integrates computational tools and cognitive-linguistic insights. By combining corpus analysis with AI-supported semi-automation, it provides a replicable and scalable model for investigating spatio-temporal metaphors, paving the way for future psycholinguistic validation and interdisciplinary applications.

References

Casasanto, D., & Boroditsky, L. (2008). Time in the mind: Using space to think about time. Cognition, 106(2), 579–593. https://doi.org/10.1016/j.cognition.2007.03.004

Cox, S., & Little, C. (2022). Time ontology in OWL [Candidate Recommendation Draft]. World Wide Web Consortium (W3C). https://www.w3.org/TR/owl-time/

Gentner, D., Imai, M., & Boroditsky, L. (2002). As time goes by: Evidence for two systems in processing space - time metaphors. Language and Cognitive Processes, 27(5), 537–565. https://doi.org/10.1080/01690960143000317

Hardie, A. (2014). CQPweb - combining power, flexibility and usability in a corpus analysis tool. International Journal of Corpus Linguistics, 17(3), 380–409. https://doi.org/10.1075/ijcl.17.3.04har

Hauser, D. J., Carter, M. S., & Meier, B. P. (2009). Mellow Monday and furious Friday: The approach-related link between anger and time representation. Cognition and Emotion, 23(6), 1166–1180. https://doi.org/10.1080/02699930802358424

Illán Castillo, R. (2024). The semantics of motion verbs within temporal conceptualization in English and Spanish: Understanding spatial dynamic construals of time [Doctoral Thesis, University of Murcia]. http://hdl.handle.net/10201/141204

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

Lakoff, G., & Johnson, M. (1980). Metaphors we live by. University of Chicago Press. https://doi.org/10.2307/430414

Loerman, A. C., & Milfont, T. L. (2018). Time after time: A short-term longitudinal examination of the ego- and time-moving representations. Journal of Research in Personality, 74, 1–5. https://doi.org/10.1016/j.jrp.2017.12.002

Margolies, S. O., & Crawford, E. L. (2008). Event valence and spatial metaphors of time. Cognition and Emotion, 22(7), 1401–1414.

McGlone, M. S., & Pfiester, R. A. (2009). Does time fly when you're having fun, or do you? Affect, agency, and Embodiment in Temporal communication. Journal of Language and Social Psychology, 28(1), 104–111. https://doi.org/10.1177/0261927X08325744

OpenAI. (2025). ChatGPT (nov 2025 version). OpenAI. https://chat.openai.com/

Richmond, J., Wilson, C., & Zinken, J. (2012). A feeling for the future: How does agency in time metaphors relate to feelings? European Journal of Social Psychology, 42(7), 813–823. https://doi.org/10.1002/ejsp.1906

Ruscher, J. B. (2011). Moving forward: The effect of spatiotemporal metaphors on perceptions about grief. Social Psychology, 42(3), 225–230. https://doi.org/10.1027/1864-9335/a000066

Soriano, C., & Piata, A. (2022). The affect bias in the metaphorical representation of anticipated events: The case of approach. Metaphor and the Social World, 12(1), 115–137. https://doi.org/10.1075/msw.18034.pia

Valenzuela Manzanares, J., & Illán Castillo, R. (2022). A corpus-based look at time metaphors. In A. Piata, A. Gordejuela, & D. Alcaraz Carrión (Eds.), Time representations in the perspective of human creativity (pp. 15–40). JohnBenjamins. https://doi.org/10.1075/hcp.75.01val

Pop Goes Profanity: Vulgarity Trends in English Chart Hits from 2000 to 2024

FULL PAPER

Lara Putensen (Universität Bonn, Germany) & Martin Schweinberger (University of Queensland, Australia)

Mainstream popular music offers a revealing lens on how language indexes social identities, stances, and cultural values. While debates about explicit lyrics are recurrent, little corpus-based research has systematically examined how vulgarity is patterned and recontextualised in mainstream music over time.

This study investigates the presence and distribution of vulgar expressions in 2,476 Billboard Year-End Hot 100 singles released between 2000 and 2024, using a corpus-based approach that combines lexical detection, artist metadata, and topic modelling. The analysis builds on previous research on vulgarity (McEnery, 2004; Schweinberger & Burrdige, 2025) and explores how the frequency and intensity of vulgarity vary across time, genre, and artist demographics, and how such usage functions as a resource for stance-taking and identity work.

Results of frequency tabulation, conditional inference trees (see Levshina, 2021), and mixed-effects logistic regressions (see Winter, 2019) show a sharp rise in both the frequency and strength of vulgar language over the 25-year period, particularly after 2015, with hip-hop emerging as the most consistently explicit genre. While male solo artists contribute the largest share of vulgar songs overall, recent years reveal higher intensity among female performers. Thematic clustering around street and gang-related discourse highlights how vulgarity becomes enregistered as a marker of authenticity, resistance, and cultural belonging. These developments reflect broader sociolinguistic processes: the colloquialisation

of public discourse, the relaxation of censorship norms, and the influence of digital and streaming cultures in reshaping linguistic markets. By tracing how mainstream music negotiates the boundaries of linguistic acceptability, the study contributes to understanding how vulgarity indexes evolving social meanings and participates in wider processes of linguistic change as well as identity and image construction.

References

McEnery, T. (2004). Swearing in English: Bad language, purity and power from 1586 to the present. Routledge.

Levshina, N. (2021). Conditional inference trees and random forests. In A practical handbook of corpus linguistics (pp. 611-643). Cham: Springer International Publishing.R Core Team. (2024). R: A language and environment for statistical computing. https://www.r-project.org/

Schweinberger; M. & Burridge, K. (2025). Vulgarity in Online Discourse around the English-Speaking World. Lingua 321: 103946. (SJR: Q1) https://doi.org/10.1016/j.lingua.2025.103946.

Winter, B. (2019). Statistics for linguists: An introduction using R. Routledge.

"Overdosed on confidence" he laments on lolloping lead single 'headlines': A corpus-based study of lyric and vocal descriptive reporting verbs in online music reviews

FULL PAPER

Karolina Ryker (University of Silesia in Katowice, Poland and Sapienza University of Rome, Italy)

Reporting verbs have been studied extensively in academic genres (Thompson & Ye, 1991; Thomas & Hawes, 1994; Eckstein et al., 2025). Consequently, existing typologies of reporting verbs are primarily grounded in academic discourse, emphasizing categories such as research, cognition, and discourse acts. Online music reviews represent an underexplored non-academic genre where reporting verbs play a key role in describing both lyrical articulation and vocal delivery. In order to address this gap, the aim of this study is to identify reporting verbs pertaining to lyric and vocal description in online album reviews. The study makes use of a specialized corpus of English Online Album Reviews compiled for the purpose of the study. The corpus encompasses 240 music reviews amounting to approximately 168 000 tokens drawn from the following time ranges 2011-2013 and 2021-2023. These are album reviews spanning 5 music genres, namely pop/R&B, rock, electronic, folk/country, and rap. The texts were drawn from three music reviewing platforms with a high number of monthly visitors according to Similarweb statistics (Pitchfork 10.8 M, Slant Magazine 343.4K, NME 7.1M as of February 2025). The research questions guiding the study are: What reporting verbs are employed by professional music critics to characterize both the articulation of lyrical content and the vocal delivery of artists? What type of evaluation do these reporting verbs convey? It is a corpus-based study combining qualitative and quantitative approaches. First, reporting verbs related to lyrics and vocals were extracted

from the corpus with the use of AntConc (Anthony, 2024) and korpusomat.eu (Kieraś et al., 2018). In particular, the p-frame s/he Vs on [X] was found to abound in reporting verbs pertaining to lyric and vocal description. It is a pattern with a personal pronoun (s/he), followed by a third-person singular verb (Vs) and the preposition 'on', introducing either a song title/type, for example "he laments on lolloping lead single 'Headlines'". Secondly, reporting verbs were tagged in QDA Miner using a typology tailored to online music reviews, adapted from academic genre frameworks by Hyland (1999) and Rawlins et al. (2024). This typology comprises the following categories: statement acts, cadence acts, and stance acts. Statements acts convey neutral lyrical output without stance, e.g., she says. Cadence acts refer to vocal texture, intensity, or flow, for example (s)he spits, howls, yawns. Stance acts encode artist's rhetorical posture or reviewer's interpretation of it, for instance boasts, proclaims, wonders, confesses. The frequency of verbs in each of the groups is reported. The differences in the distribution of reporting verbs based on the music genre (he spits in rap reviews vs. he croons in pop reviews) and the type of personal pronoun (he hollers vs. she bleats) are presented. Findings indicate that there is greater lexical variety in cadence and stance acts as opposed to statement acts, with statement acts being more repetitive. Since reporting verbs carry "evaluative potential" (Thompson & Ye, 1991, p. 369), the study sheds light on how evaluation operates in online public discourse and reflects broader digital cultures of evaluation (Jaakkola, 2024).

References

Anthony, L. (2024). AntConc (Version 4.3.1) [Computer software]. Waseda University, Japan. https://www.laurenceanthony.net/software

Eckstein, G., Rawlins, J. D., & Hanks, E. (2025). Guide to reporting verbs: Citing sources in academic writing (1st ed.). Routledge.

Hyland, K. (1999). Academic attribution: Citation and the construction of disciplinary knowledge. Applied Linguistics, 20(3), 341–367.

Jaakkola, M. (2024). Pedagogical opportunities of the review genre: Learning in cultures of evaluation (1st ed.). Routledge.

Kieraś, W., Kobyliński, Ł., & Ogrodniczuk, M. (2018). Korpusomat — a tool for creating searchable morphosyntactically tagged corpora. Computational Methods in Science and Technology, 24(1), 21–27.

Rawlins, J. D., Eckstein, G., Hanks, E., Lester, E. W., Wilde, L., & Bartholomew, R. (2024). Intentional function and frequency of reporting verbs across six disciplines: A cluster analysis. International Journal of English for Academic Purposes, 4(1), 47–71.

Thomas, S., & Hawes, T. P. (1994). Reporting verbs in medical journal articles. English for Specific Purposes, 13, 129–148.

Thompson, G., & Ye, Y. (1991). Evaluation in the reporting verbs used in academic papers. Applied Linguistics, 12(4), 365–382.

Same L1, different status? A quantitative investigation of English in Cyprus and Greece

FULL PAPER

Caroline Rak & Sarah Buschfeld (TU Dortmund University, Germany)

While World Englishes (WE) and Second Language Acquisition (SLA) research have traditionally been treated as separate research paradigms, a number of recent studies have yielded results suggesting that differences between postcolonial second-language varieties of English (ESL) and non-postcolonial learner Englishes (EFL) are not as clear-cut as long assumed and that the two disciplines should work more closely together (e.g., Buschfeld, 2020; Gilquin, 2015; Percillier, 2016). In the context of bridging the paradigm gap between WE and SLA research, the present paper quantitatively compares morphosyntactic features of postcolonial Cypriot Greek English (CyE) and non-postcolonial Greek English (GrE). The analysis aims to identify in what way linguistic differences between Cypriot Greek (CG) and Standard Modern Greek (SMG) influence the frequency of local CyE and GrE features. To this end, we investigate the following research questions:

1. Do morphosyntactic features of CyE and GrE, which are typologically equivalent and typologically different in CG and SMG, display different frequencies of non-standard realization?
2. What intra- and extralinguistic factors influence the frequency of these features?

As examples for the morphosyntactic domain, we conducted a comparative quantitative analysis of the realization of it-subjects and present perfect marking in the CEDAR corpus (Cyprus English Data Analysis and Research; Buschfeld, 2010a) and GEDAR corpus (Greek English Data Analysis and Research; Buschfeld, 2010b). For both features, local variants have been reported for CyE (Buschfeld, 2013). The corpora consist of approximately 380,000 and 130,000 words, respectively.

We present findings from both descriptive and inferential statistical analyses. The inferential statistical analysis was conducted by means of the PrInDT (Prediction and Interpretation in Decision Trees) package (Weihs & Buschfeld, 2023; version 1.0.1) in R Studio (Posit Team, 2025). This method allows for creating conditional inference trees (ctrees) with high balanced accuracies and strong predictive power, even for data sets as the current, in which the distribution of standard tokens (realized it-subjects and present perfect morphology) and non-standard tokens (zero it-subjects and local forms to mark present perfect) is fairly unbalanced. In our analysis, we focused on which extra- and intralinguistic variables (REFERENTIALITY OF SUBJECTS, PRESENT PERFECT READING, COUNTRY, AGE, GENDER, OCCUPATION, TIME SPENT ABROAD, and PLACE SPENT ABROAD) have a significant influence on the realization of either the standard or non-standard features.

These analyses have revealed significant differences in the frequencies of non-standard realizations in the two varieties, which conform to typological differences between CG and SMG. Furthermore, results show that the speakers' age is a highly significant predictor

for feature realization regardless of variety. Hence, the quantitative results indicate that L1 transfer has been at work in both varieties, independent of postcolonial background. Moreover, the statistically significant influence of AGE suggests a convergence of the two varieties in terms of status, with CyE and GrE located at a continuum between ESL and EFL status. We finally discuss what these findings imply for the research fields of SLA and WE.

References

Buschfeld, S. (2010a). Cyprus English Data Analysis and Research (CEDAR). [Data set].

Buschfeld, S. (2010b). Greek English Data Analysis and Research (GEDAR). [Data set].

Buschfeld, S. (2013). English in Cyprus or Cyprus English: An empirical investigation of variety status. John Benjamins. https://doi.org/10.1075/veaw.g46?locatt=mode:legacy

Buschfeld, S. (2020). Language acquisition and World Englishes. In D. Schreier, M. Hundt & E. W. Schneider (Eds.), Cambridge Handbook of World Englishes (pp. 559–584). Cambridge University Press.

Gilquin, G. (2015). At the interface of contact linguistics and second language acquisition research: New Englishes and Learner Englishes compared. English World-Wide, 36(1), 91–124. https://doi.org/10.1075/eww.36.1.05gil

Percillier, M. (2016). World Englishes and Second Language Acquisition: Insights from Southeast Asian Englishes. John Benjamins. https://doi.org/10.1075/veaw.g58

Posit Team (2025). RStudio: Integrated Development Environment for R(Version 2025.05.1+513) [Computer Software]. Posit Software, PBC. https://posit.co/download/rstudio-desktop/

Weihs, C., & Buschfeld, S. (2023). PrInDT: Prediction and Interpretation in Decision Trees for Classification and Regression (Version 1.0.1) [R package]. https://doi.org/10.32614/CRAN.package.PrInDT

"Why Don't You...?" Revisited: Corpus-Pragmatic Pathways Through the Jungle of Contextual Complexity

WIP

Elena Pleshakova (University of Bonn, Germany)

The construction "Why don't/doesn't/didn't NP V?" has long attracted attention in corpus pragmatics. Although its syntactic structure is stable, its illocutionary function varies, appearing as part of distinct speech acts: suggestions, requests, offers, interrogatives, etc., depending on its position in the sentence (Adolphs & Chen, 2021), the presence of communicative markers (Couper-Kuhlen, 2014), and broader contextual constellations. Variation in illocutionary act is context-dependent (Culpeper & Semino, 2000) and also shaped by "meso-level context types" (Haugh et al., 2021:6), including text genres, speech events, and activity types. These and other studies emphasize the multidimensional and dynamic origin of illocutionary acts, arising when micro-level linguistic features interact with meso-level contextual factors (Culpeper, 2021:24).
A multidimensional approach, one of the dominant trends in contemporary corpus studies

(Anthony et al., 2012), offers a framework to investigate causality and variation in illocutionary acts, exemplified by the communicative chunk "Why don't/doesn't/didn't NP V?". The use of AI-assisted methods opens new horizons for the investigation of illocutionary dynamics and causal relations in speech, which is the aim of the current exploratory study. The following hypotheses are formulated:

- The illocutionary function of the chunk "Why don't/doesn't/didn't NP V?" is a multicausal phenomenon: an interplay of micro- and meso-context variables determines its classification and allows prediction of speech-act type.
- The impact of language and socio-cultural variables on each other varies across contexts.

Drawing on over 2,000 instances of "Why don't/doesn't/didn't NP V?" extracted from the enTenTen21 corpus via SketchEngine, the study applies multidimensional coding that integrates (1) textual metadata (topic, genre, text type), (2) grammatical variables (tense of "do", R1 element, position in the sentence), and (3) pragmatic features (speech acts, pragmalinguistic strategies). The coding scheme is inspired by Adolphs (2008). By combining statistical modelling and qualitative interpretation, the study examines how minimal contextual fluctuations (e.g., changes in topic domain or clause-final markers) emerge in diverse contextual constellations to produce different illocutionary outcomes. Random-Forest and Conditional Decision-Tree modelling (Levshina, 2021) are employed not merely for prediction but also to visualize relations and causal dependencies among variables, illustrating how pragmatic meaning emerges from the interplay of multiple contextual dimensions.
Preliminary findings indicate that small contextual shifts, such as the interaction of clause-initial discourse markers with genre, systematically trigger distinct speech-act interpretations. These findings suggest that speech-act realization is multicausal and adaptive, shaped by concurrent linguistic and situational dimensions.
While enTenTen21 offers substantial coverage, it remains limited to written, web-based texts, frequently blends narrative and dialogic contexts, includes multiple varieties of English, and lacks prosodic or turn-taking information. Consequently, interpreting illocutionary force requires careful contextual and pragmatic annotation to compensate for these limitations. Nevertheless, this exploratory research demonstrates a viable step toward corpus-based investigation of dialogical data.
Overall, the study advances multidimensional pragmatic corpus research and highlights the potential of AI-assisted analysis for exploring complex communicative patterns. It shows how corpus linguistics, when combined with multidimensional coding and interpretable machine learning, can reveal the relational architecture underlying everyday communicative behavior.

References

Adolphs, S. (2008). Corpus and context: Investigating pragmatic functions in spoken discourse. John Benjamins.

Adolphs, S., Chen, Y. (2021). Corpus Pragmatics. In M. Haugh, D. Z. Kádár, & M. Terkourafi (Eds.), The Cambridge Handbook of Sociopragmatics (pp. 639 – 662). Cambridge University Press.

Anthony, L., Nishina, Y., Takahashi, K., Handford, M. (2012). Current trends in corpus linguistics: Voices from Britain (paper). Retrieved from https://laurenceanthony.net/research/20120000_english_corpus_studies/ecs_19_anthony_et_al.pdf

Couper-Kuhlen, E. (2014). What does grammar tell us about action? Pragmatics, 24(3), 623 – 647. https://doi.org/10.1075/prag.24.3.08cou

Culpeper, J., Semino, E. (2000). Constructing witches and spells: Speech acts and activity types in Early Modern England. Journal of Historical Pragmatics, 1(1), 97 – 116. https://doi.org/10.1075/jhp.1.1.08cul

Culpeper, J. (2021). Sociopragmatics: Roots and Definition. In M. Haugh, D. Z. Kádár, & M. Terkourafi (Eds.), The Cambridge Handbook of Sociopragmatics (pp. 15 – 29). Cambridge University Press.

Haugh, M., Kádár, D., Z., Terkourafi, M. (2021). Introduction: Directions in Sociopragmatics. In M. Haugh, D. Z. Kádár, & M. Terkourafi (Eds.), The Cambridge Handbook of Sociopragmatics (pp. 1 – 12). Cambridge University Press.

Levshina, N. (2021). Conditional inference trees and random forests. In M. Paquot & S. T. Gries (Eds.), A practical handbook of corpus linguistics (pp. 611 – 643). Springer. https://forms.mpi.nl/publications/item3253245/conditional-inference-trees-and-random-forests

Seek the company, and you shall find the man who keeps it: identifying semantic prosodies through measures of context valence

WIP

Mathias Russnes (University of Oslo, Norway)

This paper investigates how effectively measures of context valence can be used to identify units with distinct semantic prosodies. Semantic prosody describes how seemingly 'neutral' items recur in evaluative contexts, and is a well-known concept within corpus linguistics (Sinclair 1996; Stewart 2010). However, studies of semantic prosody have also been criticised, inter alia, for their reliance on manual analysis, because of issues of replicability and subjectivity (Dilts & Newman 2006; Bednarek 2008; Winter 2019). To tackle this issue, Snefjella & Kuperman (2016) introduced the more quantitative measure context valence, which has been adopted and modified in subsequent research (e.g. Winter 2016, 2019). In such studies, measures of context valence have been shown to effectively categorise the evaluative contexts of items, as well as to identify the impact of sense and register on semantic prosody (Author(s) in-prep). To achieve this, emotional lexicons with valence ratings were employed to calculate the mean scores of the content words in an item's context.

Building on the aforementioned studies, this paper aims to identify items without apparent evaluations that recur in negative contexts, tackling the following research question:

• Can semantic prosodies be identified through measures of context valence?

With this aim, a test study has been performed, drawing on material from the Freiburg-Brown Corpus of American English (Frown) and the Freiburg-LOB Corpus of British English (F-LOB), as well as the emotional lexicon compiled by Warriner et al. (2013). The items in this lexicon have been assigned valence scores ranging from 1 (negative) to 9 (positive), based on participant ratings. In the study, the corpora were imported into R (accessed through RStudio 3.2.1), and concordance lines were created for all instances of all items. These concordance lines consisted of five content words preceding and succeeding the node. Using the lexicon, the weighted context valence (Author(s) in-prep) of all these instances were then calculated. Following this, items where at least 30% of instances occurred within concordance lines that received weighted context valence scores lower than 4 within each corpus were extracted, following the thresholds set by Warriner et al. (2013). In addition to clear evaluative items, such as ILLNESS, TERRORIST, EPIDEMIC etc., this produced items with previously ascribed negative prosodies, such as COMMIT (Bublitz 1996) and UNDERGO (Stubbs 2001). Further, other items with less apparent negative evaluations were also identified, such as PREVENTION, INTENT, SUBSTANCE and EXEMPT, where context valence scores were found to differ markedly from the item's lexicon scores.

These preliminary results suggest that measures of context valence can identify both established and previously unrecognised semantic prosodies. The next phase will scale the method to larger corpora, such as the British National Corpus 2014, aspiring to devise a replicable method for detecting and categorising prosodies.

References

Bednarek, M. (2008). Semantic preference and semantic prosody re-examined. Corpus Linguistics and Linguistic Theory, 4(2), pp. 119-139.

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

Dilts, P. and J. Newman. 2006. 'A note on quantifying 'good' and 'bad' prosodies'. Corpus Linguistics and Linguistic Theory 2(2), pp. 233-242.

R Core Team. 2024. R: A language and environment for statistical computing. https://www.r-project.org/

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

Snefjella, B. & Kuperman, V. (2016). It's all in the delivery: effects on context valence, arousal, and concreteness on visual word processing. Cognition 156, pp. 135-146.

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

Stubbs 2001. Words and Phrases. Corpus studies of Lexical Semantics. Oxford: Blackwell Publishing.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal and dominance for 13,915 English lemmas. Behavioral Research Methods 45, pp. 1191-1207.

Winter, B. 2016. 'Taste and smell words form an affectively loaded and emotionally flexible part of the English lexicon'. Language, Cognition and Neutroscience 31(8), pp. 975-988.

Winter, B. 2019. Sensory Linguistics. John Benjamins Publishing Company.

Extraction of Multi-word Expressions in English: Combining Traditional Approaches with Information Theory

WIP

Sergei Bagdasarov, Diego Alves & Elke Teich (Saarland University, Germany)

In this study, we propose a method for automatic extraction of multi-word expressions (MWEs) that combines traditional and information-theoretic techniques. We understand MWEs as conventionalized word sequences that are to some extent processed as prefabricated units. MWEs are a very heterogeneous group, but, despite their formal and functional differences, they seem to share a common property of offering predictable transitions from one token to the other due to their high degree of conventionalization, resulting in a cognitive advantage during processing (Siyanova-Chanturia et at., 2011; Tremblay et al., 2011).

Modern systems for automatic MWE extraction often rely on deep neural networks (Ramisch et al., 2023; Savary et al., 2017) that require manually annotated datasets for training, which are not always available even for English. Moreover, neural networks typically operate as black boxes, hindering the interpretation of the outcome. In contrast, our aim is to develop a statistical extraction procedure capable of identifying MWEs of different types, striking an acceptable balance between universality and accuracy. While still relying on text corpora to compute statistical metrics, it does not require annotated data and is transparent in terms of interpretation.

Our approach combines metrics traditionally used in MWE research (frequency, association, and dispersion) with surprisal. Surprisal is an information-theoretic measure that quantifies the predictability of a word given its context (Shannon 1948). It is calculated as the negative logarithm of the conditional probability of a word. Predictable words have lower surprisal, while unexpected words show higher surprisal values. Given that each next token within an MWEs becomes more predictable, the token surprisal within a true MWE sequence should decrease.

For each MWE candidate, we compute average surprisal of each component token and then calculate both surprisal slope (overall trend in the change of token surprisal) and the surprisal delta between the last and penultimate token. Surprisal values are obtained from the smallest GPT-2 model with 124 million parameters using the surprisal Python library. As for association, we operationalize it as normalized Kullback-Leibler divergence following Gries (2022) and calculate the mean between forward and backward association for each MWE candidate. Dispersion is calculated as the proportion of texts in which an MWE candidate occurs at least once.

As a next step, we apply filters for each metric, complemented with a structural filter based on part-of-speech tags. Word sequences that successfully pass all thresholds should qualify as MWEs. Applied to the English Universal Dependencies treebank corpora, the method shows promising results, extracting MWEs of different types, for example, named entities (Buenos Aires, Harry Potter), prepositional and phrasal verbs (depend on, figure out), light verb constructions (take a look at, to make sure), collocations (highly recommended, reason-

able prices), complex function words (according to, with respect to), discourse organizers (on the other hand, in other words), and lexical bundles (I would like to, I believe that). The complete evaluation of the extraction procedure is ongoing.

References

Gries, S. Th. (2022). Multi-Word Units (and Tokenization More Generally): A Multi-Dimensional and Largely Information-Theoretic Approach. Lexis, 19. https://journals.openedition.org/lexis/6231

Kullback, S. & Leibler, R. A. (1951). On Information and Sufficiency. The Annals of Mathematical Statistics, 22(1),79–86.

Ramisch, C., Walsh, A., Blanchard, T. & Taslimipoor, Sh. (2023). A Survey of MWE Identification Experiments: The Devil is in the Details. In A. Bhatia, K. Evang, M. Garcia, et al. (Eds.), Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023) (pp. 106–120). Association for Computational Linguistics.

Savary, A., Ramisch, C., Cordeiro, S., et al. (2017). The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In S. Markantonatou, C. Ramisch, A. Savary & V. Vincze (Eds.), Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017) (pp. 31–47). Association for Computational Linguistics.

Shannon, C. E. (1948). A Mathematical Theory of Communication. The Bell System Technical Journal, 27(3), 379–423.

Siyanova-Chanturia, A., Conklin, K. & van Heuven, W. J. B. (2011). Seeing a phrase "time and again" matters: The role of phrasal frequency in the processing of multiword sequences. Journal of Experimental Psychology: Learning, Memory, and Cognition, 37(3), 776–784.

Tremblay, A., Derwing, B., Libben, G. & Westbury, Ch. (2011). Processing Advantages of Lexical Bundles: Evidence from Self-Paced Reading and Sentence Recall Tasks. Language Learning, 61, 569–613.

Corpus Methods and Software E113 • 14:00–15:30

SynFlow: Continuous Semantics Change Analysis via Dependency Co-occurences

SOFTWARE DEMO

Bach Phan Tat, Kris Heylen & Stefano De Pascale (KU Leuven, Belgium)

Modern approaches in the study of semantic change often consist of applications of vector space modelling (VSM) (Periti & Montanelli, 2024; Tahmasebi et al., 2021; Tahmasebi & Dubossarsky, 2023). However, there are many drawbacks of these methods, such as the sensitivity to corpus size (Antoniak & Mimno, 2018; Sahlgren & Lenci, 2016) and the lack of interpretability (Lenci et al., 2022). We introduce a new approach that is conceptually simpler, does not require as much data as VSM, yet is more direct in interpreting the changes in different dimensions (i.e., syntactic slots) of words' usages and meanings, along with the

corresponding package, SynFlow (Phan-Tất, 2025).

The advantages of our method are:

1. Simpler and more direct: Given a target lemma (e.g., car) and a dependency-parsed corpus, we can extract its dependency slots (e.g., adjective modifier) and slot-fillers (e.g., red) then use Jensen-Shannon Divergence (JSD) (Menéndez et al., 1997) to measure distributional changes of the slot-fillers of individual slots. This reveals how much a word has change, and in what dimensions (i.e., slots).
2. Disentangled dimensions: A conceptually related approach has been pursued by McEnery et al. (2022), who relies on surface co-occurrences. However, surface co-occurrence often suffers from accidental and/or indirect co-occurrences and the arbitrary choice of the span size (Evert, 2008), we instead adopt syntactic co-occurrence (Evert, 2008; Seretan, 2011) to separate signals.
3. Consecutive pair-wise analysis:: Although similar functions were implemented in Sketch Engine (Word Sketch Difference, 2019), it can only work with two corpora at a time so multi-period analyses require manual aggregation whereas SynFlow can do this automatically in a single pass. Moreover, while logDice (Rychlý, 2008) is well-suited for ranking collocational salience within individual slices, it does not capture profile-level dynamics. JSD over slot-filler distributions would yield a single, interpretable change estimate for each transition with an additive decomposition into per-collocate contributions that collectively sum to the total shift.
4. Small-corpus friendly: Most VSMs are not usable with sparse datasets. Finetuning pretrained models often carries information from the pretraining data (Underwood et al., 2025), making the final output unreliable. To demonstrate SynFlow's ability to work well with small corpora, we will use a subset of the Royal Society Corpus (Fischer et al., 2020), with an average of 215,000 tokens per period and a total vocab size of 70,000.

In the demonstration, we will walk the audience through the different steps and functions of the workflow (in Jupyter, with minimal programming requirement), from preparing the corpus to exploring the slots distribution and analysing the slot-fillers distributional shifts. We will also demonstrate other features (e.g., slot combinations, specialization grouping) and discuss future features of SynFlow (e.g., combination with type embeddings for slot-filler clustering). We also plan to benchmark SynFlow against other approaches with SemEval 2020's task 1 (Schlechtweg et al., 2020). Audiences are also encouraged to discuss possible research questions that SynFlow could help address.

References

Antoniak, M., & Mimno, D. (2018). Evaluating the Stability of Embedding-based Word Similarities. Transactions of the Association for Computational Linguistics, 6, 107–119. https://doi.org/10.1162/tacl_a_00008

Evert, S. (2008). Corpora and collocations. In Corpus Linguistics. An International Handbook.

Fischer, S., Knappen, J., Menzel, K., & Teich, E. (2020). The Royal Society Corpus 6.0: Providing 300+ Years of Scientific Writing for Humanistic Study.

Lenci, A., Sahlgren, M., Jeuniaux, P., Cuba Gyllensten, A., & Miliani, M. (2022). A comparative eval-

uation and analysis of three generations of Distributional Semantic Models. Language Resources and Evaluation, 56(4), 1269–1313. https://doi.org/10.1007/s10579-021-09575-z

McEnery, T., Brezina, V., & Baker, H. (2022). 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

Menéndez, M. L., Pardo, J. A., Pardo, L., & Pardo, M. C. (1997). The Jensen-Shannon divergence. Journal of the Franklin Institute, 334(2), 307–318. https://doi.org/10.1016/S0016-0032(96)00063-4

Periti, F., & Montanelli, S. (2024). Lexical Semantic Change through Large Language Models: A Survey. ACM Computing Surveys, 56(11), 1–38. https://doi.org/10.1145/3672393

Phan-Tất, B. (2025). SynFlow [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.17414457

Rychlý, P. (2008). A Lexicographer-Friendly Association Score.

Sahlgren, M., & Lenci, A. (2016). The Effects of Data Size and Frequency Range on Distributional Semantic Models. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 975–980. https://doi.org/10.18653/v1/D16-1099

Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., & Tahmasebi, N. (2020). SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. Proceedings of the Fourteenth Workshop on Semantic Evaluation, 1–23. https://doi.org/10.18653/v1/2020.semeval-1.1

Seretan, V. (2011). Syntax-Based Collocation Extraction. Springer Netherlands. https://books.google.be/books?id=I9mO0r7HbXUC

Tahmasebi, N., Borin, L., Jatowt, A., Xu, Y., & Hengchen, S. (2021). Computational approaches to semantic change.

Tahmasebi, N., & Dubossarsky, H. (2023). Computational modeling of semantic change (No. arXiv:2304.06337). arXiv. https://doi.org/10.48550/arXiv.2304.06337

Underwood, T., Nelson, L. K., & Wilkens, M. (2025). Can Language Models Represent the Past without Anachronism? (No. arXiv:2505.00030). arXiv. https://doi.org/10.48550/arXiv.2505.00030

Word sketch difference. (2019, May 14). https://www.sketchengine.eu/guide/word-sketch-difference-compare-words/

Predicting frame element coreness in FrameNet: A machine learning approach

FULL PAPER

Vladimir Buskin (Catholic University of Eichstätt-Ingolstadt, Germany)

The FrameNet project is a large-scale frame-semantic database with a seemingly usage-based core: It draws on 200,000 annotated sentences from representative corpora and offers the most comprehensive description of semantic valency patterns in English to date (Fillmore et al. 2012; Ruppenhofer et al. 2016; Boas 2020; Boas et al. 2024). Nevertheless, its empirical validity is weakened by the lack of statistical information on the distribution of lexical units, frames, and frame elements. Similarly, the characterisation of frame el-

ements as core, core-unexpressed, peripheral, or extra-thematic – intended to indicate their essentiality to a frame – is primarily motivated on theoretical grounds. For instance, verbal lexical units such as *boil*, *brown*, or *fry* are said to evoke the APPLY_HEAT frame, in which a COOK applies heat with a certain TEMPERATURE_SETTING to FOOD using a HEATING_INSTRUMENT or a CONTAINER (cf. examples 1–3).^1 While these frame elements are mostly specific to the APPLY_HEAT frame, speakers may also choose to supply circumstantial elements (e.g., MANNER or PLACE) to flexibly modify this or other events.

(1) [They COOK] boil [them FOOD] [in an iron saucepan CONTAINER].
(2) [You COOK] can brown [it FOOD] [in hot fat MEDIUM] [...].
(3) [She COOK] was frying [eggs and bacon and mushrooms FOOD] [on a camp stove HEATING_INSTRUMENT] [in Woolley's billet PLACE].

FrameNet distinguishes between different degrees of 'coreness' among the elements evoked as part of a frame. More specifically, there are core and non-core elements, with the core category subsuming core and core-unexpressed elements, and the non-core category peripheral and extra-thematic ones. For example, COOK, FOOD, HEATING_INSTRUMENT, and CONTAINER are classified as core to the APPLY_HEAT frame in the database. This raises the question of whether these labels are consistent with actual language use.

After extracting frequency data from Python's NLTK FrameNet Corpus (Bird et al. 2009) for all attested combinations of verbs, frames, and frame elements, hierarchical gradient boosting models were trained on information-theoretic measures, such as surprisal and Shannon entropy, to predict the coreness of frame elements. The performance metrics (MCC, AUC, F1, among others) reveal a nuanced picture: While some distinctions are firmly grounded in distributional patterns, others appear to be more theory-driven. The core versus non-core distinction emerges as robustly usage-based, as the models achieve near-perfect and well-balanced classification accuracy. By contrast, the rare core-unexpressed elements, which are claimed to be inherited from more abstract frames (Ruppenhofer 2016: 25), are virtually impossible to distinguish from overtly expressed core elements in terms of their distributional signatures. Similarly challenging to capture are the peripheral and especially the extra-thematic types.

These findings offer the first systematic statistical validation of FrameNet's coreness hierarchy and raise important questions for both computational lexicography and usage-based linguistics. While some theoretical distinctions find clear support in corpus distributions, others appear to require reconsideration or refinement. This work demonstrates how machine learning approaches can illuminate the empirical foundations of linguistic resources and contributes to ongoing debates about the relationship between theoretical frameworks and usage patterns in language description.

^1: Cf. `https://framenet.icsi.berkeley.edu/fnReports/data/frameIndex.xml?frame=Apply_heat`.

References

Bird, Steven, Edward Loper, & Ewan Klein. 2009. Natural Language Processing with Python. O'Reilly Media Inc.

Boas, Hans C. 2020. A roadmap towards determining the universal status of semantic frames. In Enghels, Renata, Bart Defrancq, & Marlies Jansegers (Eds.), Empirical and Methodological Challenges, 21–52. De Gruyter.

Boas, Hans C., Josef Ruppenhofer, & Collin Baker. 2024. FrameNet at 25. International Journal of Lexicography 37(3). 263–284.

Fillmore, Charles J., Russell Lee-Goldman, & Russell Rhomieux. 2012. The FrameNet Con- structicon. In Boas, Hans C. & Ivan A. Sag (Eds.), Sign-Based Construction Grammar, 193 309–372. Stanford: CSLI Publications.

Ruppenhofer, Josef, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, Collin F. Baker, & Jan Scheffczyk. 2016. FrameNet II: Extended Theory and Practice. url: https://framenet2.icsi.berkeley.edu/docs/r1.7/book.pdf.

Tracking Diachronic Lexical and Grammatical Change in Scientific English Across Disciplines

FULL PAPER

Diego Alves & Elke Teich (Saarland University, Germany)

This study analyses diachronic changes in the lexical and grammatical features of scientific English across different disciplines to determine whether trends identified in scientific English as a whole are also reflected within individual disciplines. Previous studies (e.g., Degaetano-Ortlieb and Teich, 2018; Bizzoni et al., 2020), using the Royal Society Corpus (RSC; Fischer et al., 2020), have shown that grammatical features in scientific English display a diachronic tendency toward convergence and stylistic stabilization, whereas the lexical level exhibits a slightly less accentuated convergence with a more variable pattern, reflecting ongoing innovation and adaptation to emerging scientific concepts.

The Royal Society Corpus comprises over 40,000 contributions to the Philosophical Transactions and Proceedings of the Royal Society of London from 1665 to 1996. The corpus reflects the evolution of scientific publishing from broadly interdisciplinary early issues to the later two-section format (Section A: mathematical and physical sciences; Section B: biological sciences). Thus, to examine whether similar lexical and grammatical trends occur across different scientific fields, we first classified the RSC texts by discipline using large language models (LLMs). After evaluating eight LLMs and prompting strategies on a manually annotated test subset of the RSC—balanced across historical periods and covering the disciplines of astronomy, biology, chemistry, computer science, earth sciences, engineering and technology, humanities, mathematics, medicine, miscellaneous, and physics—three models were selected for achieving the highest accuracy scores (≈0.82): gpt-oss-20b (OpenAI et al., 2025), Qwen3-8B, and Qwen3-32B-AWQ (Yang et al., 2025). The RSC texts were then annotated with these three LLMs, and we retained texts for which at least two models agreed on the label, resulting in a total of 46,988 texts (98% of the entire RSC).

To detect evolutionary trends in scientific English per discipline, we applied Kullback-Leibler Divergence (KLD; Kullback and Leibler, 1951), a measure of the difference between probability distributions. KLD quantifies how one distribution diverges from another, highlighting which linguistic features contribute most to the difference. In our study, distributions were built from lemmas (lexical level) and part-of-speech trigrams (grammatical level) within each discipline. Following Degaetano-Ortlieb and Teich (2018), we computed KLD over sliding 20-year windows with a 5-year step to identify periods of linguistic change.
Our preliminary results confirm previous findings based on the entire RSC: both the lexical and grammatical levels exhibit a general trend toward convergence, with the grammatical level showing a more pronounced effect. However, both the humanities and miscellaneous categories show trends that differ from those observed in the other fields. The miscellaneous label was assigned to texts that are not strictly scientific, such as obituaries or tables of contents, and therefore do not necessarily follow the conventions of scientific communication. At the lexical level, some disciplines show pronounced innovation peaks, leading to less standardized patterns. Astronomy, for example, peaks in lexical divergence at the late 19th century (spectrometry) and early 20th century (X-ray research). Computer science, emerging in the 1950s, starts with high divergence in part-of-speech trigrams but quickly shifts to lower values, reflecting rapid grammatical standardization.

References

Bizzoni, Y., Degaetano-Ortlieb, S., Fankhauser, P., & Teich, E. (2020). Linguistic variation and change in 250 years of English scientific writing: A data-driven approach. Frontiers in Artificial Intelligence, 3, 73.

Degaetano-Ortlieb, S., & Teich, E. (2018). Using relative entropy for detection and analysis of periods of diachronic linguistic change. Saarländische Universitäts-und Landesbibliothek.

Fischer, S., Knappen, J., Menzel, K., & Teich, E. (2020, May). The Royal Society Corpus 6.0: Providing 300+ years of scientific writing for humanistic study. In Proceedings of the Twelfth Language Resources and Evaluation Conference (pp. 794-802).

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. The annals of mathematical statistics, 22(1), 79-86.

OpenAI, S. A., Lama Ahmad, J. A., Sam Altman, A. A., Edwin Arbus, R. K. A., & Yu Bai, B. B. (2025). gpt-oss-120b & gpt-oss-20b Model Card. arXiv preprint arXiv:2508.10925.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., ... & Qiu, Z. (2025). Qwen3 technical report. arXiv preprint arXiv:2505.09388.

Strategies of Onomastic Referencing in 18th-Century British Grammar Writing

FULL PAPER

Beatrix Busse, Nina Dumrukcic & Sophie Du Bois (University of Cologne, Germany)

Beyond being essential components of a classical education, grammar books of the Early

Modern English period held a "traditional authority" and "the right to make language decisions" (Mitchell 1994, 548-549). Especially in the 18th century, which is frequently termed the age of prescriptivism (see Locher 2008, p.130; Auer 2009), proclamations of what constitutes 'correct' language use, also known as the 'doctrine of correctness' (Leonard 1929/1962) were common in grammar writing.

For these reasons, grammar writing in the 18th century has been investigated from a range of different perspectives. For instance, in-depth analyses of individual prescriptivist authors such as Robert Lowth (e.g. Tieken-Boon van Ostade 2010) and Lindley Murray (e.g. Fens de Zeeuw 2011) have shed light on the role of grammars at the time.

On a wider scope, within the HeidelGram project, investigations of British grammar writing from the 16th, 17th, and 19th centuries have indicated changing and stable norms over time. For instance, shifts in not only who is being referenced in these grammar texts, but also which strategies are employed to do so were analysed. In a previous investigation of 18th-century grammars more than 6000 onomastic references were extracted and categorized for the types of persons being referred to, such as political figure and literary author (see Busse et al. 2025). The study has shown the onset of the influence of the major prescriptivist Robert Lowth. Furthermore, the notable rise in references to poets and literary authors has indicated the lasting effects of the English Renaissance on English grammar writing.

This paper aims to extend these findings by investigating the strategies employed by the grammar authors when referring to other persons. Additional information of the reference strategies employed when referring to major prescriptivists such as Lowth, or influential literary persons such as William Shakespeare and John Milton, will provide a more holistic view on how the grammar authors of the 18th century positioned themselves in regard to these persons. Comparisons to findings from the previous centuries will enable the detection of trends in diachrony.

The 18th-century component of the HeidelGram corpus (Busse et al. 2015–) will serve as a basis for this study. It comprises a total of 1,5 million tokens taken from 24 grammar books. For each extracted reference, the person referred to is identified and categorized and the reference strategy employed is assigned. There are six reference categories, such as opinion and quotation, which were originally established for the 19th-century grammar data (see Busse et al. 2018, 2019, 2020).

Due to the rise in prescriptive writing at the time, we expect to find an increase in the use of opinions as compared to previous centuries. Additionally, since 18th-century grammarians were said to "rely heavily on each other's work" (Locher 2009, p. 131), we expect to see significant intercitation between the grammars using quotations, acknowledgements, and comparisons. We will compare findings from the 18th century with our previous analyses of the most salient types of references in the 17th and 16th century strands of the corpus.

References

Auer, A. (2009). The Subjunctive in the Age of Prescriptivism. Palgrave Macmillan UK. https://doi.org/10.1057/9780230584365

Busse, B., Kleiber, I., Dumrukcic, N., & Du Bois, S. (2025). Onomastic Referencing in 18th-Century

British Grammar Writing. ICAME46, Vilnius, Lithuania.

Busse, B., Gather, K., & Kleiber, I. (2020). A Corpus-Based Analysis of Grammarians' References in 19th-Century British Grammars. In A. Cermakova & M. Malá (Eds.), Diskursmuster - Discourse Patterns: Vol. 20. Variation in Time and Space: Observing the World Through Corpora. De Gruyter.

Busse, B., Gather, K., & Kleiber, I. (2019). Paradigm Shifts in 19th-Century British Grammar Writing: A Network of Texts and Authors. In B. Bös & C. Claridge (Eds.), Norms and Conventions in the History of English. John Benjamins.

Busse, B., Gather, K., & Kleiber, I. (2018). Assessing the Connections between English Grammarians of the Nineteenth Century: A Corpus-Based Network Analysis. In Eric Fuß, Marek Konopka, Beata Trawiński, & Ulrich H. Waßner (Eds.), Grammar and Corpora 2016 (pp. 435–442). Heidelberg University Publishing.

Busse, B., Gather, K., & Kleiber, I. (2015–). HeidelGram. A Corpus of English Grammar Books between 1550 and 1900.

Fens-de Zeeuw, L. (2011). Lindley Murray (1745-1826), Quaker and grammarian. LOT: Vol. 283. LOT, Netherlands Graduate School of Linguistics.

Leonard, S. A. (1962). The Doctrine of Correctness in English Usage 1700-1800. Russel & Russel Inc. (Original work published 1929)

Locher, M. A. (2008). Chapter 7: The Rise of Prescriptive Grammars on English in the 18th Century. In J. A. Fishman, M. A. Locher, & J. Strässler (Eds.), Contributions to the Sociology of Language. Standards and Norms in the English Language (Vol. 95, pp. 127–148). Mouton de Gruyter. https://doi.org/10.1515/9783110206982.1.127

Mitchell, L. C. (1994). Inversion of grammar books and dictionaries in the seventeenth and eighteenth centuries. In Proceedings of the 7th International Conference on Lexicography and Grammar (pp. 546–559). Amsterdam: Euralex.

Tieken-Boon van Ostade, I. (2010). The Bishop's Grammar: Robert Lowth and the Rise of Prescriptivism. Oxford University Press.

White, H. D. (2011). Scientific and Scholarly Networks. In J. Scott & P. Carrington (Eds.), The SAGE Handbook of Social Network Analysis (pp. 271–285). SAGE. https://doi.org/10.4135/9781446294413.n19

The Hidden Politics of Historical American Grammar Writing

FULL PAPER

Sophie Du Bois (University of Cologne, Germany)

Historical grammars of English have been said to convey a sense of national identity, although this has received limited attention thus far. For instance, Mitchell (2012, 2020) has found anti-foreigner sentiment in 17th- and 18th-century writings on language, while Wolf (2010) has identified discourses of language and nation in 18th-century British grammars of English. Du Bois (forth.) has investigated the multifaceted linguistic choices with which American grammarians of the 18th and 19th centuries contributed to the creation and

maintenance of an explicitly American identity. Using a mixed method approach, three strategies for conveying an American national identity were identified, namely 1) onomastic referencing, 2) branding America as a nation, and 3) recalling a shared heritage (Du Bois forth.).

This paper explores, how natural language processing (NLP) approaches might extend these findings. More specifically, it was observed in the above study that the grammar authors express a multitude of opinions within each of the three strategies. However, the extraction of these opinions was limited by the methodology to the extracted onomastic references and search terms.

In this study, large language models (LLMs) from the huggingface (Wolf et al. 2019) library are employed to perform sentiment analyses on the Historical American Grammar Corpus (HistAGram) (Du Bois forth.), which is a 1-million-word corpus of 18th- and 19th-century American grammar texts. The aim is to extract positive or negative opinions more generally. For this purpose, a sentiment analysis pipeline is built using Python. Different models, fine-tuned to the sentiment analysis task, are utilized and their performance evaluated.

Two main obstacles are identified in this process. Firstly, historical data is notoriously difficult to process in an automated manner due to historical language use, bad print quality, irregular spelling, and many more factors (Claridge 2008). Additionally, most language models are trained on modern English data, which further limits usability.

Secondly, the textbook nature of the works is particularly obstructive in this task as the texts contain generic example sentences, which can be assigned positive and negative sentiments, although they do not necessarily reflect the authors' positions. Nonetheless, these sentences are included in the analyses, since it was previously indicated that at least some of these example sentences seem to be deliberately chosen by the grammarians (Du Bois forth.).

Methodologically, the analyses indicate that the choice of LLM plays a critical role in the quality of the findings. For instance, while some models assign values of negative or positive sentiment, other models additionally include a neutral sentiment.

The findings indicate that, although the texts were written during the age of prescriptivism (Locher 2008, Auer 2009), about 80% of sentences seem to display a neutral sentiment, while only the remaining 20% are either positive or negative. And although strong opinions can be found in the texts, there is no clear tendency towards more positive or negative utterances over time.

References

Auer, A. (2009). The Subjunctive in the Age of Prescriptivism. Palgrave Macmillan UK. https://doi.org/10.1057/9780230584365

Claridge, C. (2008). Historical corpora. In A. Lüdeling & M. Kytö (Eds.), Handbücher zur Sprach- und Kommunikationswissenschaft: Vol. 29.1. Corpus Linguistics: An International Handbook (pp. 242–259). Berlin: De Gruyter.

Du Bois, S. (forthcoming). The Hidden Politics of Grammar: Unveiling Discourses of American Identity in a Corpus of Historical American Grammar Writing. Applied Corpus Linguistics. Rout-

ledge.

Locher, M. A. (2008). Chapter 7: The Rise of Prescriptive Grammars on English in the 18th Century. In J. A. Fishman, M. A. Locher, & J. Strässler (Eds.), Contributions to the Sociology of Language. Standards and Norms in the English Language (Vol. 95, pp. 127–148). Mouton de Gruyter. https://doi.org/10.1515/9783110206982.1.127

Mitchell, L. C. (2012). Language and National Identity in 17th- and 18th-century England. In C. Percy & M. C. Davidson (Eds.), Multilingual Matters: v.148. The Languages of Nation: Attitudes and Norms (1st ed., pp. 123–140). Channel View Publications.

Mitchell, L. C. (2020). Grammar Wars: Seventeenth - and Eighteenth - Century England. In C. L. Nelson, Z. G. Proshina, & D. R. Davis (Eds.), Blackwell Handbooks in Linguistics. The Handbook of world Englishes (Second edition, pp. 475–494). Wiley Blackwell.

Wolf, G. (2010). The 'Language of the bravest, wisest, most powerful, and respectable Body of People': The Discourse on Language and Nation in 18th-Century Grammars of English. In C. Lange, U. Schaefer, & G. Wolf (Eds.), Linguistics, ideology and the discourse of linguistic nationalism: Workshop held during the inaugural conference of International Society for the Linguistics of English (ISLE 1) from 8th to 11th October 2008 in Freiburg, Germany (pp. 53–76). Lang.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., Platen, P. von, Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q. Rush, A. M. (2019, October 9). HuggingFace's Transformers: State-of-the-art Natural Language Processing. https://arxiv.org/pdf/1910.03771

Improving automated transcription to compile large and diverse historical corpora

FULL PAPER

Tanja Säily (University of Helsinki, Finland), Ari Vesalainen (University of Helsinki, Finland), Samuli Kaislaniemi (University of Eastern Finland, Finland), David Denison (University of Manchester, UK) , Nuria Yáñez-Bouza (University of Vigo, Spain), Pete Morris (University of Manchester, UK), Akseli Kettunen (University of Helsinki, Finland) & Inga Kokkonen (University of Helsinki, Finland)

Transcription has traditionally been expert work requiring a great deal of time and resources. Therefore, corpora transcribed from manuscripts tend to be small, which limits their use to the study of relatively high-frequency phenomena. While print sources are easier to access and digitize, relying on them alone is problematic, as they tend to represent a very small section of society and more formal language use. By contrast, manuscript sources like personal letters are more 'speech-like' (Culpeper & Kytö 2010: 17) and could be written by anyone who was literate.

There is, then, a need for corpora that are both large and socially diverse, the compilation of which will require more automated methods. Recent developments in large language models show promise for many tasks related to corpus development (e.g. Säily et al. 2025).

The AI-based Transkribus tool now provides access to massive 'super models' that work fairly well for transcribing standard data. However, our experiments indicate that these perform poorly on the letters of lower-class, less educated writers, who are underrepresented in the training data.

The Corpus of Early English Correspondence Extension is a socially representative dataset covering the long eighteenth century (c.1680–1800; Kaislaniemi 2018). Based on published editions of letters ranging from the working-class Clift family to Georgian royalty, the CEECE is among the largest existing corpora of correspondence (2.2Mw), though still dwarfed by corpora of Late Modern English print texts like CLMET and COHA. Yet the CEECE can be used as a stepping-stone for the next generation of corpora of historical correspondence, for which purpose we have photographed over 700 CEECE letters in archives throughout the UK.

We are developing a text recognition model specialized for 18th-century handwritten materials by leveraging recent advances in vision-language models (VLMs). Rather than relying on architectures such as TrOCR (Li et al., 2023), we adopt models from the Qwen family, which have demonstrated strong performance in OCR-related tasks (Vesalainen et al. 2026). Training data will be drawn from multiple historical datasets, including the CEECE, The Mary Hamilton Papers, and David Hume's manuscripts, in order to capture linguistic and stylistic variation across different social contexts. This combined dataset enables the development of a more generalizable model of 18th-century handwriting that spans diverse writers and registers. The resulting model will be evaluated against state-of-the-art Transkribus models and other VLM-based approaches to assess both transcription accuracy and cross-domain generalization, highlighting the Qwen-based model's strengths in these areas.

The new model will be made publicly available, and we expect it to significantly facilitate the compilation of larger and more diverse Late Modern English corpora. To extend its coverage, the model can be trained further with existing 19th-century hand-transcribed corpora and manuscript images (e.g. Auer et al. 2024). Moreover, the workflow developed in our project will enable the research community at large to train models for other languages and time periods, benefiting not only linguists but also historians and other scholars working in the digital humanities.

References

Auer, A., Gardner, A., & Iten, M. (2024). Creating a corpus of Late Modern English pauper letters: Uncertainties, challenges, and solutions. Swiss Papers in English Language and Literature, 44, 121–140.

CEECE = Corpus of Early English Correspondence Extension. Compiled by T. Nevalainen, H. Raumolin-Brunberg, S. Kaislaniemi, M. Laitinen, M. Nevala, A. Nurmi, M. Palander-Collin, T. Säily, & A. Sairio at the Department of Languages, University of Helsinki. https://varieng.helsinki.fi/CoRD/corpora/CEEC/

CLMET = The Corpus of Late Modern English Texts, version 3.1. Compiled by H. De Smet, S. Flach, H.-J. Diller, & J. Tyrkkö. https://fedora.clarin-d.uni-saarland.de/clmet/clmet.html

COHA = Davies, M. (2010–). The Corpus of Historical American English: 400 million words, 1810–2009. https://www.english-corpora.org/coha/

Culpeper, J., & Kytö, M. (2010). Early Modern English dialogues: Spoken interaction as writing. Cambridge University Press.

Kaislaniemi, S. (2018). The Corpus of Early English Correspondence Extension (CEECE). In T. Nevalainen, M. Palander-Collin, & T. Säily (Eds.), Patterns of change in 18th-century English: A sociolinguistic approach (pp. 45–59). John Benjamins.

Li, M., Lv, T., Chen, J., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z., & Wei, F. (2023). TrOCR: Transformer-based optical character recognition with pre-trained models. Proceedings of the AAAI Conference on Artificial Intelligence, 37(11), 13094–13102. https://doi.org/10.1609/aaai.v37i11.26538

Säily, T., Suomela, J., Perek, F., Jiménez Real, J., & Vartiainen, T. (2025). Using large language models to enrich corpus metadata: The case of novels in COHA. Paper presented at the 46th Annual Conference of the International Computer Archive of Modern and Medieval English (ICAME 46), Vilnius, Lithuania, June 2025. https://tanjasaily.fi/talks/icame46_saily_et_al_2025.pdf

The Mary Hamilton Papers (c.1740–c.1850). Compiled by D. Denison, N. Yáñez-Bouza, T. Oudesluijs, C. Ulph, C. Wallis, H. Barker, & S. Coulombeau, University of Manchester, 2019–2023. https://doi.org/10.48420/21687809

Vesalainen, A., Mäkelä, E., Ruotsalainen, L., & Tolonen, M. (2026). Error patterns in historical OCR: A comparative analysis of TrOCR and a vision-language model. arXiv preprint arXiv:2602.14524.

Characterising Late Modern English Historiographical Writing via Situation Entity Types

FULL PAPER

Claudia Claridge, Hanna Schmück & Annemarie Friedrich (Universität Augsburg, Germany)

How did 18th to early 20th century authors narrate history? We approach this question via annotating, computationally modelling and analysing linguistic features related to discourse modes (Smith, 2003), integrating corpus linguistics with state-of-the-art artificial intelligence methods. Discourse modes such as NARRATIVE, DESCRIPTION, REPORT, INFORMATION and ARGUMENT represent text passages that are characterised by clusters of linguistic features and help to describe progression within texts and genre differences. The linguistic features include Situation Entities (SEs), which roughly correspond to aspectual clause types and include EVENTS, STATES, GENERIC SENTENCES (statements about "kinds"), or GENERALIZING SENTENCES (generalisation over situations), QUESTIONS, and IMPERATIVES.

We focus on Late Modern English historiographical writing to quantify how authors differ in their use of SEs as a proxy for discourse mode (Palmer and Friedrich, 2014). Automatic classifiers for SE types (Dai & Huang, 2018; Friedrich et al., 2016; Rezaee et al., 2021) have been developed based on the SitEnt dataset (Friedrich et al., 2016), which contains contemporary English. There is, to the knowledge of the authors, no method or benchmark for SE annotation and analysis of historical varieties of English. We present a new gold standard dataset of 859 manual SE annotations using the annotation scheme by Friedrich and Palmer (2014). The data represents historiographies written by seven different authors between 1702 and 1909 taken from the much larger (1.5M words, 50 authors) corpus of

Late Modern English History Writing (CLMEH, Claridge [forthcoming]). Annotation of further data is ongoing.

We further contribute a traditional corpus-linguistic analysis of differences in SE type distributions of different authors. Our findings show strong idiosyncratic differences: Marsden, for example, presents a larger share of GENERIC SENTENCES (33.8% vs the average of 8.6%) in combination with a lower share of EVENTS 5.4% vs the average of 38.4%) than the remaining sampled authors which indicates a comparative lack of narrative elements in his writing.

To scale SE annotation in historical English corpora we benchmarked three computational models for SE classification. We evaluated the zero-shot performance of the large language model Llama-3.3-70B-Instruct and fine-tuned the transformer encoders XLM-RoBERTa (Conneau et al., 2020) and MacBERTh (Manjavacas & Fonteyn, 2021) on the SitEnt dataset. We achieved the following accuracies and macro-F1 scores for the four main SE types EVENTS, STATES, GENERIC and GENERALIZING SENTENCES on our historiography gold dataset:

Model Accuracy Macro-F1
Llama-3.3-70B-Instruct 0.5496 0.3815
XLM-RoBERTa 0.6024 0.4570
MacBERTh 0.6526 0.4670

All models most robustly identify EVENTS and STATES. The two transformer encoders fail on GENERALIZING SENTENCES and perform better on GENERIC SENTENCES while the opposite pattern emerges for Llama. Overall results are promising, yet not approaching literature results for contemporary English, possibly due to the linguistic variability and genre-specific features of Late Modern English historiography. MacBERTh, trained on historical English data, outperformed the other models, indicating that pretraining on genre- and period-appropriate corpora significantly enhances classification accuracy. These findings highlight the need for further fine-tuning on manually annotated historical data. The best-performing classifiers will, given satisfactory performance, be deployed to automatically annotate the full CLMEH.

References

Claridge, C. (forthcoming, 2026). History Writing. In M. Kytö & E. Smitterberg (Eds.), The New Cambridge History of the English Language: Documentation, Sources of Data and Modelling (Vol. 2, pp. 433–458). Cambridge University Press.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 8440–8451.

Dai, Z., & Huang, R. (2018). Building Context-aware Clause Representations for Situation Entity Type Classification. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 3305–3315.

Friedrich, A., & Palmer, A. (2014). Situation entity annotation. Proceedings of the 8th Linguistic Annotation Workshop (LAW VIII) (pp. 149–158).

Friedrich, A., Palmer, A., & Pinkal, M. (2016). Situation entity types: Automatic classification of

clause-level aspect. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 1757–1768. https://doi.org/10.18653/v1/P16-1166

Manjavacas, E., & Fonteyn, L. (2021). MacBERTh: Development and Evaluation of a Historically Pre-trained Language Model for English (1450-1950). Proceedings of the Workshop on Natural Language Processing for Digital Humanities (NLP4DH) Co-Located with ICON-2021, 23–36. https://doi.org/10.26615/978-952-94-5833-2_004

Palmer, A., & Friedrich, A. (2014). Genre distinctions and discourse modes: Text types differ in their situation type distributions. Proceedings of the Symposium on Frontiers and Connections between Argumentation Mining and Natural Language Processing.

Rezaee, M., Darvish, K., Kebe, G. Y., & Ferraro, F. (2021). Discriminative and Generative Transformer-based Models For Situation Entity Classification. CoRR, abs/2109.07434.

Schweter, S. (2024). Pretrained Language Models on British Library Corpus (Version 1.0.0) [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.10715629

Smith, C. (2003). The modes of discourse. The local structure of texts. Cambridge University Press.

Highs and Lows: Modeling Ambivalence in Online Cannabis Discourse

FULL PAPER

Andrew Lustig (University of Toronto, Canada) & David Chartash (Yale University, New Haven, CT, USA)

This paper investigates how ambivalence—the coexistence of conflicting positive and negative attitudes—manifests linguistically in online cannabis discourse. Ambivalence lies at the emotional core of addiction and recovery, capturing the tension between approach and avoidance, pleasure and harm. Yet beyond the clinical domain, it also functions as a discursive resource through which users negotiate identity, belonging, and moral stance. This study asks: how is ambivalence linguistically enacted and distributed across communities that variously celebrate, moderate, or renounce cannabis use?

The data comprise the 1,000 most recent discussion threads from three Reddit communities representing distinct orientations toward cannabis: r/trees (advocacy and celebration), r/petioles (moderation and harm reduction), and r/leaves (cessation and recovery). These corpora were collected using the Python Reddit API Wrapper (PRAW) and cleaned for analysis, yielding approximately 8 million tokens of user-generated text. Reddit was chosen for its scale, pseudonymity, and structured community boundaries, which make it an ideal site for examining how subcultural discourse reflects and reshapes affective norms around substance use.

Our approach combines traditional corpus linguistic methods—keyword and collocation analysis—with an AI-based ambivalence classifier. Using a few-shot prompt with GPT-4-

turbo, each comment was rated on a continuous scale from 0 (no ambivalence) to 1 (high ambivalence). This operationalizes ambivalence as a measurable linguistic affect: a pattern of contradiction and self-address through which emotion is enacted in language. In this sense, ambivalence can be read as a constellation of affective intensities—fluctuations of feeling that course through discourse and cluster around moments of hesitation, irony, or self-contradiction.

Results reveal distinct affective profiles aligned with community ethos. r/trees shows overwhelmingly positive affect, marked by insider humour and ironic excess ("So good it will literally kill you"), suggesting playful awareness of overuse. r/leaves and r/petioles display comparable levels of high ambivalence ("I don't want to—I do want to—I see the harm it does"), reflecting shared struggles with moderation and self-control. Differences between these two communities are minimal, but both diverge sharply from r/trees.

These findings suggest that ambivalence is a pervasive emotional signature of recovery and moderation discourse online. The paper argues that large language models can illuminate the fine-grained emotional dynamics of addiction talk, transforming qualitative intuitions about tension and uncertainty into quantifiable linguistic signals. By linking corpus linguistics with digital psychiatry, Highs and Lows demonstrates how corpus research in the age of AI can trace the affective textures of change and self-reflection in digital publics.

Beyond the cannabis context, this study models how linguistic ambivalence functions as a broader marker of conflicted self-relation in digital talk about substance use, mood, and identity. In tracing these affective tensions computationally, the project illustrates how corpus research in the age of AI can bridge linguistic patterning and psychosocial meaning, yielding insights relevant to both discourse analysts and mental-health researchers.

References

Baker, P. (2006). Using corpora in discourse analysis. Continuum.

Brookes G. Insulin restriction, medicalisation and the Internet. Communication & Medicine. 2019 Jul 3;15(1):14-27.

McGlashan, M., & Krendel, A. (2023). Complement keywords: Corpus comparison and distinctiveness in online discourse. International Journal of Corpus Linguistics, 28(3), 411–435.

Miller, W. R., & Rollnick, S. (2012). Motivational interviewing: Helping people change (3rd ed.). Guilford Press.

OpenAI. (2025). GPT-4-turbo API documentation. OpenAI.

Plural Identities and the Hermeneutics of Suspicion: A Corpus-Assisted Analysis of Online Multiplicity

FULL PAPER

Andrew Lustig (University of Toronto, Canada & Centre for Addiction and Mental Health), Oliver Harrison (University of Toronto, Canada), Mark McGlashan (University of Liverpool), Gavin Brookes (Lancaster University), Sophie Simic-Lustig (Hudson College) & Benoit Mulsant (University of

Toronto, Canada)

This paper explores how plural communities on Reddit construct, contest, and negotiate legitimacy in relation to psychiatric and cultural discourses. Plurality—also known as multiplicity—refers to the experience of multiple selves or "parts" within a single body. While historically framed as pathological within psychiatry, plural systems increasingly articulate their identities outside diagnostic frameworks, often redefining multiplicity as volitional, creative, or spiritual. Online platforms such as Reddit have become key sites for these negotiations, where suspicion and empathy circulate as competing hermeneutic logics.

The study analyzes four Reddit communities that anchor different positions within the plural spectrum: r/DID, r/OSDD, r/tulpas, and r/plural. Using a 1.3-million-word corpus drawn from the top 200 most-upvoted threads in each subreddit, we apply a mixed-methods approach combining keyword analysis, set-theoretic comparison, and qualitative concordance reading. "Complement keywords"—lexical items distinctive to one community but absent from others—serve as entry points for interpreting how legitimacy is linguistically performed. These keywords were generated using WebCorp and analyzed in Sketch Engine, producing a lexicon of 467 distinctive items across communities.

Findings reveal distinct discursive configurations of legitimacy and affect. r/DID relies on authority-claiming, invoking diagnostic language and professional expertise while simultaneously critiquing psychiatry's limits. r/OSDD is saturated with suspicion and gatekeeping terms such as claimers, impostor, and pretending, reflecting what Fricker (2007) terms testimonial injustice. r/tulpas, by contrast, constructs legitimacy through world-building and practice—keywords such as wonderland, thoughtform, and art index creative hermeneutics of empathy largely devoid of suspicion. Finally, r/plural centers on boundary-realignment, with terms like endo, non-disordered, and inclusive signaling efforts to build an explicitly broad and affirming counterpublic.

Across communities, the hermeneutics of suspicion and empathy (Ricoeur, 1970) map onto different orientations toward psychiatric authority, trauma, and volition. Plural discourse on Reddit thus exemplifies what Fraser (1990) and Warner (2002) call counterpublics: self-organizing discursive spaces where marginalized groups negotiate recognition within algorithmic publics. By identifying linguistic markers of epistemic injustice and hermeneutic repair, the study shows how plural systems generate new interpretive resources—blurry, wonderland, endo—that make contested experiences intelligible on their own terms.

Beyond its theoretical contributions, this analysis demonstrates how corpus-assisted discourse methods can illuminate the linguistic infrastructures of belief, care, and legitimacy in digital mental-health publics. In doing so, it argues that corpus research, when combined with interpretive frameworks from philosophy and psychiatry, can help capture the affective and epistemic textures of online identity work—an inquiry increasingly vital in the age of AI-mediated discourse.

References

Fraser, N. (1990). Rethinking the Public Sphere: A Contribution to the Critique of Actually Existing Democracy. Social Text, no. 25/26 (1990): 56–80.

Fricker, M. (2007). Epistemic injustice: Power and the ethics of knowing. Oxford University Press.

McGlashan, M., & Krendel, A. (2024). Keywords of the manosphere. International Journal of Corpus Linguistics, 29(1), 87–115.

Ricoeur, P. (1970). Freud and philosophy an essay on interpretation. Yale University Press.

Warner, M. Publics and Counterpublics. New York: Zone Books, 2002.

Out with Corpora, in with ChatGPT & Co.? English Students' Attitudes Towards Corpora and AI as Correction Aids for Academic Writing

FULL PAPER

Katharina Deckert (University of Bamberg, Germany)

Various studies have deemed a corpus-based approach to error correction effective and students judge it as beneficial but generally demanding (e.g. Bridle, 2019; Crosthwaite, 2017) since extensive practice is necessary for students to become corpus literate (Schlüter & Deckert, 2025). At the University of Bamberg, Germany, a data-driven learning (DDL) seminar aimed at fostering English students' corpus literacy for applied purposes with a focus on error detection and correction has been taught and evaluated since 2020 (Großmann & Schlüter, 2024).

Recently, easily accessible AI applications, which can be employed to automatically correct texts quite successfully (Schlüter & Deckert, 2025), have arisen as a potential competition for traditional corpus literacy and DDL (Crosthwaite & Baisa, 2023). Despite their advantages, AI tools have limitations as well (Barrot, 2023), which is why it is of the utmost importance to foster students' AI literacy (Ng et al., 2021) and raise awareness of the fact that "[c]ombining corpus-based DDL with GenAI appears to be 'a useful methodological synergy'" (Crosthwaite & Baisa, 2023, p. 100066).

Therefore, the aforementioned seminar was rebooted in 2024. In the new course design students acquire theoretical and practical knowledge of corpora and AI before learning how to apply both to correct typical learner errors in the English language. The final session of the seminar invites students to come to their own informed conclusions regarding the initial question: "Out with Corpora, in with ChatGPT & Co.?".

This contribution will present quantitative and qualitative results of analyses that were conducted over the course of three semesters with the aim of finding out about (a) students' approaches to (academic) writing, (b) their referencing habits and preferences, and (c) their attitudes towards corpora and AI tools as writing aids as well as how these attitudes changed over the course of the seminar.

Preliminary qualitative analyses of pre-test (N=24+) and post-test (N=14+) questionnaires, which were based on previous research (Quinn, 2015; Yoon & Hirvela, 2004), and individual

interviews (N=17+) during the first two semesters suggest that hardly any of the students had used a corpus before taking the seminar, while most were already familiar with AI tools. After the seminar, students had different tendencies regarding which of the two they preferred depending on the type of text they had to write as well as the type of language problem in question. Students were aware of the main advantages and disadvantages of corpora and AI tools and many concluded that using a combination of the two might be the most beneficial approach: "If you're writing a term paper, use both ChatGPT and corpus. One is not enough." (Interviewee5, 23:26). Finally, the analyses show that after having been exposed to both corpora and AI tools for an entire semester, most students anticipated that they would keep using AI tools for text correction, but in a more critical and reflected way than before – indicating that they had indeed become more AI literate –, and they had learned to appreciate corpora as an additional and reliable reference.

References

Barrot, J. S. (2023). Using ChatGPT for second language writing: Pitfalls and potentials. Assessing Writing, 57, 100745. https://doi.org/10.1016/j.asw.2023.100745

Bridle, M. (2019). Learner use of a corpus as a reference tool in error correction: Factors influencing consultation and success. Journal of English for Academic Purposes, 37, 52–69. https://doi.org/10.1016/j.jeap.2018.11.003

Crosthwaite, P. (2017). Retesting the limits of data-driven learning: Feedback and error correction. Computer Assisted Language Learning, 30(6), 447–473. https://doi.org/10.1080/09588221.2017.1312462

Crosthwaite, P., & Baisa, V. (2023). Generative AI and the end of corpus-assisted data-driven learning? Not so fast! Applied Corpus Linguistics, 3(3), 100066. https://doi.org/10.1016/j.acorp.2023.100066

Großmann, C., & Schlüter, J. (2024). Corpus Literacy in der Lehrer*innenbildung: Englisch (lernen) lehren mit Korpora. In A. Rosen & K. Beuter (Eds.), Englische Sprachwissenschaft und Fachdidaktik im Dialog: Chancen zur Stärkung der Lehrkräftebildung (pp. 185–202). Narr Francke Attempto Verlag. https://doi.org/10.24053/9783381112524

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. Computers and Education: Artificial Intelligence, 2, 100041. https://doi.org/10.1016/j.caeai.2021.100041

Quinn, C. (2015). Training L2 writers to reference corpora as a self-correction tool. ELT Journal, 69(2), 165–177. https://doi.org/10.1093/elt/ccu062

Schlüter, J., & Deckert, K. (2025). Artificial Intelligence vs. Corpus Literacy: Ansätze zur Vermittlung reflektierter Schreibkompetenz in der Wissenschaftssprache Englisch. In L. Mrohs, J. Franz, D. Herrmann, K. Lindner, & T. Staake (Eds.), Digitales Lehren und Lernen an der Hochschule: Strategien – Bedingungen – Umsetzung (pp. 111–119). transcript Verlag. https://doi.org/10.1515/9783839471203-011

Yoon, H., & Hirvela, A. (2004). ESL student attitudes toward corpus use in L2 writing. Journal of Second Language Writing, 13(4), 257–283. https://doi.org/10.1016/j.jslw.2004.06.002

Writing accuracy, IQ and motivation: A conjunction of learner corpus research and psychometric data

FULL PAPER

Lea Bracke (University of Bamberg, Germany)

As a relatively young research discipline, Learner Corpus Research (LCR) has already established itself as a valuable subdiscipline of corpus linguistics (Granger et al. 2015: 1-3). The Corpus of Young German Learner English (YGLE; Werner et al. in preparation) expands on the still underrepresented demographic of beginner to intermediate learners, and contains multimodal, pseudo-longitudinal data from grades 5 to 12. The project conceptualizes learner proficiency as an interplay of complexity, accuracy, and fluency (CAF), a triad well established in the field (Thewissen 2021; Bui and Skehan 2018; Housen et al. 2012).

Accuracy, i.e., "the ability to produce error-free speech" (Housen & Kuiken 2009: 461), has established its pivotal place within LCR as an important prerequisite for successful communication (Thewissen 2021: 311; Evans et al. 2014: 34). While accuracy has been researched extensively in both SLA and learner corpus studies (e.g. Thewissen 2021; Kowal 2018; Housen et al. 2012), insights into individual (psychological) factors that may influence accuracy are rare in LCR, with a study by Möller (2017) being a notable exception. Hence, the YGLE corpus offers an opportunity to examine the relationship between learner corpus data and psychometric variables.

The present study investigates the following research question: Which individual variables correlate with global accuracy? In this context, a particular focus lies on two psychometric variables: cognitive abilities, measured by the AID-G (Kubinger & Hagenmüller 2019), and achievement motivation, assessed by the FLM (consisting of five subscales; Petermann & Achtergarde 2015; Petermann & Lobeck 2019). Global accuracy is operationalized as errors per words, as this type of measure is easily applicable to the language of young learners who may still produce incomplete sentence structures.

For this study, a sample of N ≈ 700 learner essay was analyzed with regard to accuracy. In addition to the psychometric variables, metadata on learners' age, gender, L1 and school type were collected. The relationships between variables were examined by calculating ANOVAs as well as single- and multiple-regression modelling in RStudio.

Preliminary results indicate a stronger correlation between accuracy and cognitive abilities ($r = -0.25^{***}$) than between accuracy and achievement motivation ($r = -0.08^{*}$, and $r = -0.06$; representing two subscales). Moreover, cognitive abilities and one motivation subscale (Strive for Achievement) interact in predicting accuracy. Regarding the relationship between school type, accuracy and psychometric variables, two main trends emerge: 1) The relationship between motivation and accuracy is most pronounced in the school type "Realschule"; 2) When controlling for the influence of school type, cognitive abilities are solely a significant factor in the school type "Mittelschule". These findings suggest that in certain environments, contextual factors may outweigh psychological factors, while in other cases, learners' traits may significantly influence their acquisition of accuracy.

References

Bui, Gavin & Peter Skehan. 2018. Complexity, accuracy, and fluency. The TESOL encyclopedia of English language teaching. 1-7. https://doi.org/10.1002/9781118784235.eelt0046

Evans, Norman W., K. James Hartshorn, Troy L. Cox & Teresa M. De Jel. 2014. Measuring written linguistic accuracy with weighted clause ratios: A question of validity. Journal of Second Language Writing 24. 33-50. https://doi.org/10.1016/j.jslw.2014.02.005

Granger, Sylviane, Gaëtanelle Gilquin & Fanny Meunier. 2015. Introduction: learner corpus research – past, present and future. In: Sylviane Granger et al. (eds.), The Cambridge handbook of learner corpus research. Cambridge University Press. 1-5. https://doi.org/10.1017/CBO9781139649414

Housen, Alex & Folkert Kuiken. 2009. Complexity, accuracy, and fluency in second language acquisition. Applied linguistics 30(4). 461-473. https://doi.org/10.1093/applin/amp048

Housen, Alex, Folkert Kuiken & Ineke Vedder. 2012. Complexity, accuracy and fluency: Definitions, measurement and research. In: Alex Housen, Folkert Kuiken & Ineke Vedder (eds), Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA. John Benjamins Publishing Company. 1-20. https://ebookcentral.proquest.com/lib/ub-bamberg/detail.action?docID=1040791

Kowal, Iwona. 2018. The dynamics of complexity, accuracy and fluency in second language development. Jagiellonian University Press. https://www.cambridge.org/core/product/identifier/9788323394747/type/BOOK

Kubinger, Klaus D. & Bettina Hagenmüller. 2019. AID-G – Gruppentest zur Erfassung der Intelligenz auf Basis des AID. Testzentrale.

Möller, Verena. 2017. Language acquisition in CLIL and non-CLIL settings: learner corpus and experimental evidence on passive constructions. John Benjamins. https://ebookcentral.proquest.com/lib/ub-bamberg/detail.action?docID=5181677

Petermann, Franz & Sandra Achtergarde. 2015. FLM 7-13 - Fragebogen zur Leistungsmotivation für Schüler der 7. bis 13. Klasse, 2nd ed. Testzentrale.

Petermann, Franz & Annette Lohbeck. 2019. FLM 3-6 R – Fragebogen zur Leistungsmotivation für Schülerinnen und Schüler der 3. bis 6. Klasse – Revision. Testzentrale.

Thewissen, Jennifer. 2021. Accuracy. In: Nicole Tracy-Ventura & Magali Paquot (eds.), The Routledge Handbook of second language acquisition and corpora. Routledge. 305-317. https://ebookcentral.proquest.com/lib/ub-bamberg/detail.action?docID=6403462

Werner, Valentin, Robert Fuchs, Anna Rosen, Lyudmila Kruhlenko, Bethany Stoddard & Lea Bracke. In preparation. Corpus of Young German Learner English (unpublished).

From Concordances to Play: Gamifying Data-Driven Learning for Young ESL Learners

WIP

Cansu Akan (Chemnitz University of Technology, Germany)

Data-Driven Learning (DDL), grounded in the principle of engaging learners directly with authentic linguistic evidence (Johns, 2002), has long been recognised as a pedagogical

approach capable of fostering inductive noticing, pattern recognition, and learner autonomy. Yet despite its theoretical promise, DDL has remained predominantly associated with advanced or adult learners due to the perceived cognitive demands of interpreting concordance lines and the limitations of traditional corpus interfaces (Boulton, 2009). Consequently, its integration into primary school contexts is still rare, with research on young learners described as "very much underexplored" (Crosthwaite & Baisa, 2023). Existing work offers only preliminary insights into feasibility and learning outcomes (Pérez-Paredes, 2019; Hirata, 2019; Gatto, 2019). The development of an innovative, young-learner-friendly corpus tool therefore holds considerable potential to make linguistic data accessible, deflect fears, and bring authenticity and autonomy into the early language-learning process.

Responding to this gap, the present study aims to make corpus use pedagogically viable for learners aged 8–10 by developing an accessible multi-modal corpus application and examining the features such an environment should include. More broadly, the research seeks to evaluate the linguistic benefits and pedagogical feasibility of structured, guided exposure to concordances for young learners, with a particular focus on vocabulary development, noticing, and early rule formation. The application is conceptualised as a serious game, drawing on Zyda's (2005) definition of serious games as digital environments designed primarily for education rather than entertainment. A gamified, mobile approach to DDL provides an appealing pathway for implementation in young learner settings, offering an experience that is seamless, personal, portable, and aligned with the expectations and learning habits of Generation Alpha.

The methodological design centres on the development and evaluation of a multi-modal serious game application built on a carefully compiled pedagogical corpus of more than one million words. This corpus draws from over 300 graded readers and picture books written for CEFR A1–A2+ learners and selected according to readability metrics, lexical profiles, topic familiarity, and age appropriateness. It incorporates audio recordings, visuals, and haptic interaction tools to support auditory, visual, and kinaesthetic learning styles. Within the application, 150 DDL tasks employ colour-coded concordances to highlight collocations and colligations, increasing gradually in linguistic complexity.

The empirical component consists of a three-month intervention with 54 primary school learners in Germany. Quantitative data include pre- and post-test measures of learning gains alongside fine-grained interaction logs capturing time on task, task completion patterns, and error types. Qualitative data are collected through semi-structured learner interviews. These datasets are triangulated to produce a detailed understanding of how young learners interact with corpus evidence through corpus-informed learning materials, how they respond to gamified DDL environments, and how these interactions relate to emerging linguistic competence.

The project is expected to make substantive contributions to applied corpus linguistics, CALL, and instructional design by generating empirical evidence on DDL in primary education, establishing design principles for young-learner-oriented corpus materials, and offering a scalable model for multi-modal, corpus-driven serious game environments with the potential for future AI-supported adaptive learning.

References

Boulton, A. (2009). Data-driven learning: Reasonable fears and rational reassurance. Indian Journal of Applied Linguistics, 35(1), 81-106.

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

Gatto, M. (2019). Query complexity and query refinement: Using Web search from a corpus perspective with digital natives. In Crosthwaite, P. (Ed.). Data-Driven Learning for the Next Generation: Corpora and DDL for Pre-tertiary Learners (1st ed). Chapter 7.

Hirata, E. (2019). The development of a multimodal corpus tool for young EFL learners: A case study on the integration of DDL in teacher education. In Crosthwaite, P. (Ed.). Data-Driven Learning for the Next Generation: Corpora and DDL for Pre-tertiary Learners (1st ed). Chapter 6.

Johns, T. 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.

Pérez-Paredes, P. (2019). The pedagogic advantage of teenage corpora for secondary school learners. In Crosthwaite, P. (Ed.). Data-Driven Learning for the Next Generation: Corpora and DDL for Pre-tertiary Learners (1st ed). Chapter 5.

Zyda, M. (2005). From visual simulation to virtual reality to games. Computer, 38(9), 25–32. doi: 10.1109/MC.2005.297.

Building the Witchpaper Corpus database: towards a multidisciplinary resource for historical and linguistic research, and beyond

WIP

Christa Schneider (University of Bern, Switzerland)

The past decade has seen an increasing interest in digitising and systematically analysing early modern judicial and administrative texts. Among these, witch trial records represent a uniquely valuable yet often rather underexplored source for understanding linguistic, social, and historical practices across time and regions. This paper presents ongoing work towards the creation of a Witch Paper Database (WCD) - a digital infrastructure project designed to integrate, standardise, and enrich multiple existing, digital and analogue corpora of witch trial documents from Scotland, Switzerland, colonial North America and many more.
The WCD will combine several regional sub-corpora, including collections from Edinburgh and its vicinity, Bern and other Swiss archives, and, hopefully, also the Salem witchtrials papers. While individual collections have been partially digitised and studied, they remain scattered and heterogeneous in transcription standards, metadata formats, and accessibil-

ity. The aim of this project is to establish a unified and extensible database that facilitates both linguistic and interdisciplinary research on witch trial discourse.
The database will include standard bibliographic and contextual metadata (e.g., date, location, case type, archival reference), digitised facsimiles and transcriptions (where archival guidelines permit), possibly enriched with linguistic and structural annotation. In addition, automatically generated corpus statistics and extracted information (e.g., named entities, sentiment patterns, semantic domains) will be made visible through an interactive web interface. The long-term goal is to create a platform that not only supports linguistic and corpus-based research but also invites collaboration across history, legal studies, gender studies, and digital humanities.
This short paper reports on the conceptual design and current development stage of the project, highlighting lessons learned from previous experiments with Named Entity Recognition (NER), Sentiment Analysis (SA) and Part-of-Speech tagging (POS) on the witchtrial data. It discusses challenges related to standardisation, multilingualism and the ethical and archival constraints of digitising and sharing judicial records. A particular focus is placed on identifying the core requirements for a sustainable research infrastructure, including interoperability, FAIR data principles, and long-term preservation strategies.
The presentation aims to invite discussion and feedback from corpus and infrastructure specialists to shape the design of the database before a possible funding proposal (SNF/ERC). By presenting this as a work in progress, the paper seeks to foster collaboration and to outline a roadmap for developing a comprehensive, open, and methodologically transparent resource that will support future research well beyond linguistics.

References

Profeta, G., Rinaldi, F., & Cornelius, J. (2025, September 24). Mini-Muse final report. https://doi.org/10.17605/OSF.IO/4VT92
https://aris.supsi.ch/entities/publication/20d75c89-0254-46c5-94e3-e607c16936a4
https://mini-muse.github.io/project/

Investigating financial communication: Introducing the Earnings Calls Corpus (ECC500)

WIP

Christian Langerfeld & Gisle Andersen (Norwegian School of Economics)

Earnings calls (ECs) are an essential form of financial communication that serves to connect the management of corporations with its stakeholders. ECs are a professional and institutionalized communicative genre by which corporate managers announce their companies' quarterly or yearly results and engage in discussions with market analysts. Such calls serve as a key site for transparency and interaction between managers and financial market actors, such as investors and analysts. It gives the management an opportunity to frame the discourse surrounding financial results while also being held accountable, thus

often characterised by an 'interplay of consensus and tension' (Hirsti et al. 2023; see also Camiciottoli 2017; Nagengast 2024; Guo 2025).

In this work-in-progress presentation we describe our initiative to explore this genre from a discourse-analytical perspective using corpus linguistic methods. We describe the compilation and analytic potential of a new corpus containing the full transcripts of quarterly ECs for each firm listed in the S&P 500. The corpus covers the period 2006–2025, which includes global events such as the financial crisis and the COVID-19 pandemic. The final corpus will include approx. 35,000 transcripts. In the corpus construction procedure we capture metadata such as the name of the firm, the date of the EC, the industry the firm operates in according to the Fama/French 49 classification (Fama & French, n.d.), names and gender of the participants, and the affiliation of the analysts that are present and ask questions.

Earnings calls are a structured communicative event, typically beginning with an operator's introduction and safe-harbour disclaimer, followed by prepared remarks from senior management (usually the CEO and CFO) presenting financial results, business updates and forward-looking guidance. This is usually followed by a free-form question-and-answer (Q&A) session with financial analysts (Hynes et al., 2024; Matsumoto et al., 2011).

Each utterance is coded for speaker role and enriched with metadata, and this enables scalable corpus-linguistic analyses of pragmatic and interactional features and comparison between speakers, firms and industries. In our presentation we will make some preliminary observations about the discourse-analytic features of the EC genre and briefly outline our plans for research based on the ECC500. The corpus approach enables systematic and large-scale studies of how EC participants use persuasive language strategically to achieve their professional objectives as providers of information (executives) and compare this to the language used by discerning or challenging seekers of information (analysts). This creates a dialogic dynamic of tension and consensus, as observed by Hirsti et al. (2023), which deserves scrutiny in such a wide set of data. Particular attention is paid to hedging and politeness strategies, and forms that suggest rapport-building and role-boundary marking. The inclusion of speaker gender allows us to explore whether question formulation, response style, hedging behaviour, or politeness choices exhibit systematic gender-linked patterns (de Jong et al., 2025). We present preliminary descriptive results on structural variation and outline further possibilities for research, including diachronic patterns and cross-sector comparison.

References

Crawford Camiciottoli, B. (2017). Persuasion in Earnings Calls: A Diachronic Pragmalinguistic Analysis. International Journal of Business Communication, 55(3), 275-292. https://doi.org/10.1177/2329488417735644

Fama, E. F., & French, K. R. (n.d.). Fama/French 49 industry classification [Data set]. Kenneth R. French Data Library. Retrieved October 31, 2025, from https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

Guo, M., & Lo, K. (2025). Formal and informal language in earnings conference calls. SSRN.

https://doi.org/10.2139/ssrn.5174884

Hirsto, H., Koskela, M., & Jokipii, A. (2023). Performing financial communication as professional practice: The interplay of consensus and tension in earnings calls. Journal of Professions and Organization, 10(2), 165–181. https://doi.org/10.1093/jpo/joad011

Hynes, L., Garvey, J., & O'Brien, F. (2024). The anatomy of an earnings call. SSRN. https://doi.org/10.2139/ssrn.4696562

de Jong, P., Timmerman, I., & Doey, B. (2025). Linguistic complexity and gender in financial analysis: Evidence from earnings call questioning patterns. Journal of Behavioral Finance, 1–13. https://doi.org/10.1080/15427560.2025.2556660

Matsumoto, D., Pronk, M., & Roelofsen, E. (2011). What makes conference calls useful? The information content of managers' presentations and analysts' discussion sessions. The Accounting Review, 86(4), 1383–1414. https://doi.org/10.2308/accr-10034

Nagengast, M. (2024). Causal reasoning in earnings releases: Insights from a European perspective. SSRN. https://doi.org/10.2139/ssrn.5012425

Developing a Spoken Corpus of Engineering Lab Talk for ESP Instruction

WIP

Frederick Dunn (Purdue University, United States of America)

As technology continues to advance, so too do the ways we can design and deliver English for Specific Purposes (ESP) courses. This project focuses on the development of a specialized spoken corpus drawn from Electrical and Computer Engineering (ECE) laboratory sessions at a large tier 1 research university in the Midwestern United States. The goal of this corpus is to provide targeted language support for English language learners who will work as Teaching Assistants (TAs) in ECE labs, where communication involves procedures, directives, and problem-solving language (Tapper, 1994; Axelson & Madden, 1994; Gorsuch, 2006). Previous engineering corpora, including those related to ECE, have been based primarily on textual data such as academic documents, reports, and textbooks (Chen,2010; Ward, 2007; Hyland, 2008). However, the researcher's previous experience working in labs and with engineering students surfaced the idea that spoken language in laboratories differs substantially from what appears in text-based sources. For example, written materials often feature full technical terms such as operational amplifier (Del Toro, 1986); however preliminary findings from lab recordings show that abbreviated forms like op-amp are used in place of the full technical name in spoken interactions. This distinction highlights the need for a corpus that captures the authentic, real-time vocabulary and phraseology used by instructors and students when operating in a lab.

To better prepare International Teaching Assistants (ITAs) for the communicative demands of laboratory teaching, it is essential that language instruction reflect the realities of lab discourse. The specialized corpus described here represents the first stage of a larger initiative to integrate corpus-informed ESP instruction with virtual reality (VR) training modules

that simulate authentic ECE laboratory contexts.
This research project addresses the following research questions:
1. What directive and question types are most frequent in ECE lab talk?
2. What are the high-frequency technical terms that appear and what are the common abbreviated forms of these technical terms?
3. What are common collocations that co-occur with technical actions involving lab equipment?
Data are collected from ECE lab sessions using audio recordings of current TAs instructing mid-level labs for undergraduate classes of 4-15 students. AI-assisted transcription tools, such as Clipchamp and the university's in-house large language model (LLM) are being used to generate transcripts and annotate the transcripts, followed by manual correction focusing on directives, questions, high-frequency vocabulary. The data analyses will examine question types, frequency of collocations, and the type-token frequency of technical terms, using SketchEngine as the primary tool for corpus exploration. These linguistic findings will inform the design of communication activities and VR-based training that help future ECE ITAs use clear, accurate, and field-specific English in real laboratory environments.

References

Axelson, E. R., & Madden, C. G. (1994). Discourse strategies for ITAs across instructional contexts. Discourse and performance of international teaching assistants, 153-185.

Chen, L. (2010). An investigation of lexical bundles in ESP textbooks and electrical engineering introductory textbooks. Perspectives on formulaic language: Acquisition and communication, 107-125.

Del Toro, V. 1986 Electrical engineering fundamental. Englewood Cliffs: Prentice Hall

Gorsuch, G. J. (2006). Discipline-specific practica for international teaching assistants. English for Specific Purposes, 25(1), 90–108. https://doi.org/10.1016/j.esp.2005.06.003

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. English for specific purposes, 27(1), 4-21.

Tapper, J. (1994). Directives used in college laboratory oral discourse. English for Specific Purposes, 13(3), 205-222.

Ward, J. (2007). Collocation and technicality in EAP engineering. Journal of English for Academic Purposes, 6(1), 18-35.

The Influence of Sociolinguistic Constellations on Variation in Northeast Indian English

WIP

Tjorven Halves (University of Bonn, Germany)

Globalization and increasing international communication have made investigating variation in world Englishes highly relevant (e.g., Siemund, 2023). However, existing research tends to adopt either qualitative approaches with limited generalizability or quantitative, corpus-

based methods with little to no speaker metadata (e.g., Gries & Bernaisch, 2016; Isingoma, 2021; Sabaté-Dalmau, 2018; Szmrecsanyi & Grafmiller, 2023). This study addresses this gap by integrating quantitative methods with comprehensive sociodemographic data.

Northeast India is an under-researched region that is significantly more linguistically diverse than mainland India (Fuchs et al., 2025). English plays an important role in the region as a medium of instruction and lingua franca, especially because Hindi is not as widely used (Fuchs et al., 2025). This study examines the role of English in relation to other languages in Northeast India by statistically modeling dimensions of sociodemographic variation and dominant language constellations (DLCs; e.g., Lo Bianco & Aronin, 2020) to determine their effects on linguistic variation in a corpus of spoken English. It asks three research questions:

1) Which dimensions of sociodemographic variation exist in the population?
2) Do distinct DLC types and roles of English in them exist in the population?
3) Do the identified dimensions and DLC types affect variation in spoken English morphosyntactic complexity?

Data will be collected as part of the project "English as a local lingua franca in the multilingual ecology of Northeast India", funded by the German Research Foundation (research unit FOR 5728). A stratified random sample of Northeast Indian participants (n = 180) will complete a researcher-administered questionnaire covering sociodemographics, education, language repertoire, proficiency, use, and attitudes. A subset of informants (n = 60) will additionally complete semi-structured interviews of approximately two hours duration each, forming a corpus of spoken Northeast Indian English.

First, dimensions of sociodemographic variation will be modeled with a multiple correspondence analysis (MCA) on the questionnaire dataset. Second, DLC types will be identified by analyzing questionnaire items related to self-reported language proficiency, use, and choice with a clustering method. Third, morphosyntactic complexity in the interview corpus will be analyzed using mixed-effects regression to test whether the MCA-identified sociodemographic dimensions and the DLC groups predict linguistic variation.

The analysis is expected to reveal several sociodemographic dimensions, including for instance core demographics (e.g., age, gender) and educational trajectory (e.g., schooling, profession), and two- to four-language DLCs involving varying centrality of English. These dimensions and DLCs are anticipated to predict variation in spoken English.

In sum, the study contributes to the world Englishes and variationist sociolinguistics fields by combining quantitative corpus analysis with extensive sociolinguistic data from an underrepresented region. It examines the local English variety not in isolation, but as intertwined with its multilingual ecology, and proposes a novel, quantitative approach to modeling DLCs. Since Northeast India is representative of many postcolonial, multilingual contexts around the world, this study offers a replicable approach for investigating variation in diverse English varieties. The proposed work-in-progress report focuses on the first part of the analysis.

References

Fuchs, R., Wiltshire, C., & Sarmah, P. (2025). The role of English in the linguistic ecology of Northeast

India. In P. Siemund, G. Stein, & M. Vida-Mannl (Eds.), World Englishes in their local multilingual ecologies (Vol. 9, pp. 291–316). John Benjamins. https://doi.org/10.1075/hsld.9.13fuc

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

Isingoma, B. (2021). The sociolinguistic profile of English at the grassroots level: A comparison of Northern and Western Uganda. In C. Meierkord & E. W. Schneider (Eds.), World Englishes as the grassroots (pp. 47–69). Edinburgh University Press.

Lo Bianco, J., & Aronin, L. (2020). Introduction: the dominant language constellations: a new perspective on multilingualism. In Dominant language constellations: A new perspective on multilingualism (pp. 1–15). Springer International Publishing.

Sabaté-Dalmau, M. (2018). 'I speak small': Unequal Englishes and transnational identities among Ghanaian migrants. International Journal of Multilingualism, 15(4), 365–382. https://doi.org/10.1080/14790718.2018.1428329

Siemund, P. (2023). Multilingual development: English in a global context. Cambridge University Press.

Szmrecsanyi, B., & Grafmiller, J. (2023). Comparative variation analysis: Grammatical alternations in world Englishes. Cambridge University Press.

Referring to non-binary people as *they*: A corpus-based alternation studies on specific gendered pronouns

FULL PAPER

Helena Hanneder & Linnea Garlepow (Philipps-Universität Marburg, Germany)

Pronouns refer to notional gender, namely socially constructed groups of men, women and non-binary people. They can reflect said gender, presume gender (when using *he* or *she* even though a referent's gender is unknown) or – purposefully or not – be misused for affective reasons (McConell-Ginet 2013). In recent years, usage of the pronoun *they* has been the topic of research and political debate, since *they* has acquired a dual meaning: it can refer to anyone regardless of their gender or refer to an individual with a non-binary gender identity (Renström 2025). In this study, we will investigate the latter: the current usage of pronouns for non-binary referents.

Recent studies have shown that neo-pronouns (e.g. *ze*) are functionally similar, however, *they* was experimentally shown to be most widely understood (Renström 2025, Bradley et al. 2019), and also most favoured by non-binary individuals (Loureiro-Porto and Ariza-Fernandéz 2025). We will therefore focus on the pronoun *they* within multiple possible chosen pronoun combinations: *they/them* (referring to both a person assigned female and male at birth) and the rolling pronouns *he/they*, *she/they*, and *he/she/they*. According to a community-driven wikipage (Nonbinary Wiki 08.09.2025), the celebrities who use these pronoun combinations and who were most often mentioned in the 2025 section of the NOW

corpus were chosen for this study (Davies 2016-2025). These are Sam Smith (*they/them*, amab, 103 articles), Courtney Stodden (*they/them*, afab, 20 articles), Janelle Monáe (*she/they*, 57 articles), Elliot Page (*he/they*, 108 articles) and Jonathan Van Ness (*he/she/they*, 59 articles). We will view pronoun fidelity as an alternation that is affected by extra- and intralinguistic factors and analyse it as the response variable of a multifactorial model. It is likely that political affiliation with regard to LGBT allyship will have an effect on whether the preferred pronoun is used (Conrod 2019), which is why we will capture the political stance of the newspaper. Intralinguistically, we will annotate pronoun choice, pronoun persistency (Arnold 2025), referent topicality (Kehler & Rohde 2013), distance to referent and potential competitors (Arnold et al. 2022). Thus, we ask the questions: What are the pronouns with which non-binary celebrities are referred to, do they match their preferred pronouns and what affects pronoun alternation.

Preliminary results reveal that Elliot Page and Janelle Monáe, although using rolling pronouns, are mostly referred to with the gendered pronouns *he* or *she*, respectively. Usage of *they* and *he* for Jonathan Van Ness is almost balanced, but there is no instance of *she*, only when her preferred pronouns are listed. Sam Smith's exclusive use of the pronoun *they* is mostly respected, whereas Courtney Stodden is mostly referred to with *she/her*, rarely with *they/them*. These first patterns highlight prominent differences between the referents, which promise interesting results when considering the intra- and interlinguistic variables.

References

Arnold, J. E., Marquez, A., Li, J., & Franck, G. (2022). Does nonbinary they inherit the binary pronoun production system? Glossa Psycholinguistics, 1(1). https://doi.org/10.5070/g601183.

Arnold, J. E. (2025). Hearing Pronouns Primes Speakers to Use Pronouns. Open Mind, 9, 47–69. https://doi.org/10.1162/opmi_a_00178

Conrod, K. (2019). Pronouns raising and emerging [Doctoral dissertation, University of Washington]. University of Washington Libraries. https://digital.lib.washington.edu/researchworks/handle/1773/44673.

Davies, Mark. (2016-2025) Corpus of News on the Web (NOW). Available online at https://www.english-corpora.org/now/.

Loureiro-Porto, L., & Ariza-Fernández, J. L. (2025). Nonbinary pronouns in X (Twitter) bios: Gender and identity in online spaces. Research in Corpus Linguistics, 13(1), 171–196. Asociación Española de Lingüística de Corpus (AELINCO).

McConnell-Ginet, S. (2013). Gender and its relation to sex: the myth of 'natural' gender. In The expression of gender (pp. 3–38). De Gruyter Mouton Berlin

Nonbinary Wiki. (08.09.2025). Notable nonbinary people. https://nonbinary.wiki/wiki/Notable_nonbinary_people

Renström, E. A. (2025). The implementation of neo- and nonbinary pronouns: Areview of current research and future challenges. Frontiers in Psychology, 15, Article 1507858.

Automating Modality: AI-Driven Semantic analysis of Scottish Standard English

FULL PAPER

Johannes Trüdinger (University of Bayreuth, Germany)

Research on Scottish Standard English (SSE) traditionally takes place within the confines of the Scottish English continuum (Aitken 1979) and has a strong phonological bias (Schützler 2024). The assumption was that since SSE functions as the standard pole in the Scottish English continuum, its grammatical system is equal to that of Standard English English (SEE; Stuart-Smith 2008: 48). This, paired with a lack of appropriate tools and resources, has led to a deficit in grammatical research on SSE. SSE has since been established as a fully-fledged standard variety (Schützler 2015; Schützler et al. 2017). Yet the lack of grammatical research persists. Schützler & Herzky (2021) and Schützler (2024) led first efforts to address this issue and find differences in the use of modals of strong obligation, indicating a grammatical system different, at least in part, from SEE.

The present study expands on this previous research, conducting an analysis of the core modal verbs (can, could, may, might, shall, should, will, would) in spoken SSE and two references varieties – Standard American English (SAE) and Standard English English (SEE). It asks: (i) whether SSE exhibits distinct frequency profiles of core modals relative to SAE and SEE; (ii) whether the distribution of modality senses in Palmer's (2001) framework (epistemic/deontic/dynamic) differs across the varieties; and (iii) whether a fully automated random forest based classifier can reliably assign modality sense at scale.

The dataset comprises 38,000 modal tokens drawn from International Corpus of English-Scotland (Schützler et al. 2017), the British National Corpus (BNC 2014), and the Santa Barbara Corpus of Spoken American English (Du Bois 2000-2005). Pre-processing uses a fine-tuned BART (Lewis et al. 2020) model for punctuation restoration and sentence segmentation and the frame semantic transformer (Chanin 2023). Model validation used an 80/20 train-test split on manually annotated data. The best model currently achieves 89% accuracy, with further improvement expected as the training data is expanded.

For frequency, we compare normalized rates across corpora. For modality sense distributions, we fit a multinomial model with modality sense as the response and corpus as the predictor. Preliminary results indicate that SSE differs not only in modal frequencies but also in the distribution of modality senses, with significant deviations from both SAE and SSE. These findings provide new evidence for grammatical distinctiveness of SSE and demonstrate that automated, large-scale semantic analysis of modality is possible.

References

Aitken, A. J. (1979). Scottish Speech: a historical view with special reference to the StandardEnglish of Scotland. In: Aitken, A. J.; McArthur, Tom (eds.), Languages of Scotland.Edinburgh: W & R

Chambers. 85-119.

Chanin, David (2023). Open-Source Frame Semantic Parsing. arXiv:2303.12788.https://arxiv.org/abs/2303.12788.

Du Bois, John W., Wallace L. Chafe, Charles Meyer, Sandra A. Thompson, RobertEnglebretson, and Nii Martey (2000-2005). Santa Barbara corpus of spoken AmericanEnglish, Parts 1-4. Philadelphia: Linguistic Data Consortium.

Lewis, Mike; Liu, Yinhan; Goyal, Naman; Ghazvininejad, Marjan; Mohamed, Abdelrahman;Omer, Levy; Stoyanov, Veselin; Zettlemoyer, Luke (2020). Bart: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, andComprehension. In: Proceedings of the 58th Annual Meeting of the Association forComputational Linguistics. 7871-7880. Online: Association for ComputationalLinguistics. https://aclanthology.org/2020.acl-main.703/.

Love, R., Dembry, C., Hardie, A., Brezina, V. and McEnery, T. (2017). The Spoken BNC2014: designing and building a spoken corpus of everyday conversations. In: International Journal of Corpus Linguistics, 22(3), pp. 319-344.

Palmer, Frank R. (2001). Mood and Modality. Cambridge: Cambridge University Press.

Schützler, Ole (2015). A Sociophonetic Approach to Scottish Standard English. AmsterdamJohn Benjamins.

Schützler, Ole (2024). The Elusive Butterfly of Scottish Standard English. In: ChristineElsweiler (ed.), The languages of Scotland and Ulster in a global context, past andpresent. Selected Papers from the 13th triennial Forum for Research on the Languagesof Scotland and Ulster, Munich 2021 (Publications of FRLSU 8). Aberdeen: FRLSU.91-138).

Schützler, Ole; Gut, Ulrike; Fuchs, Robert (2017). New perspectives on Scottish StandardEnglish: Introducing the Scottish component of the International Corpus of English. In:Hancil, Sylvie; Beal, Joan C. (eds.), Perspectives on Northern Englishes. Boston MA:De Gruyter Mouton. 274-300.

Schützler, Ole; Herzky, Jenny (2021). Modal verbs of strong obligation in Scottish StandardEnglish. In: English Language & Linguistics 26 (1). 133-159.

Stuart-Smith, Jane (2008). Scottish English: phonology. In: Kortmann, Bernd; Upton, Clive(eds.), Varieties of English 1: The British Isles. Berlin: Mouton de Gruyter. 48-70.

The *Digital Lexical Atlas of Scotland*: The Issue of 'what is a word?' and word frequency

FULL PAPER

John Kirk (University of Vienna, Austria)

The background to this paper is the *Digital Lexical Atlas of Scotland* (DLAS) (Kirk, in preparation), a collaborative project at the University of Vienna and the Austrian Academy of Sciences. It is based on the data of the original *Linguistic Atlas of Scotland* (Mather & Speitel 1975, 1977).

The data are representative of the traditional folk vocabulary of Scots, which ranges from words for human beings, their bodies, their clothes, their characteristics, to children's games, the natural world, including insects, beasts and farm animals, and to the land and traditional (usually manual) ways of farming the land and animal husbandry – all referring to concepts

which have been in oral currency for centuries.
Structurally, DLAS is based on a relational database, which is linked to an interactive mapping program. But in view of the author's long experience with corpora, they cannot but conceive of the data as a corpus. But it is a corpus with several differences: as it comprises only lexical phenomena, every word counts as significant.
Except that is the issue: whereas the original data amounts to nothing but ortho-phonological variants, what actually is to be counted as a word? When word frequency is considered, to what do the frequencies actually refer? The outcome of critical ortho-phonological data analysis is the establishment of lexical word types – being called here lexemes. How many lexemes are there in this corpus? Moreover, how far can a lexeme have variants, and on what ground are these bring established? The critical interrogation of the DLAS data builds on McArthur (1999) and Kirk (2009) to identify the 11 possible categorisations of words which have been informing the establishment of lexemes. Only after the lexemes are established can the maps be drawn, for the discovery of frequencies and the cartographical display of geographical distributions.

References

Kirk, John M. 2009. 'Word Frequency: Use or Misuse?' in Dawn Archer (ed.) *What's in a Word List? Investigating Word Frequency and Keyword Extraction*. London: Ashgate. pp. 17–34.

Kirk, John, Markus Pluschkovits, Hans Christian Breuer, and Ludwig Maximilian Breuer. In preparation. *The Digital Lexical Atlas of Scotland*. University of Vienna & Austrian Academy of Sciences. https://lasdb.dioe.at/ [accessed: 15 May 2026].

Mather, J. Y., and Speitel, H.-H. (eds.) 1975. *The Linguistic Atlas of Scotland*: Scots section, vol. 1. London: Croom Helm.

Mather, J. Y., and Speitel, H.-H. (eds.) 1977. *The Linguistic Atlas of Scotland*: Scots section, vol. 2. London: Croom Helm.

McArthur, Tom. 1999. 'The word "word"'. In his *Living Words: Language, Lexicography and the Knowledge Revolution*. Exeter: Exeter University Press, 42–49.

The new Corpus of Patient Documents-UK: coverage – case studies – potential

FULL PAPER

Julian Mader (LMU Munich, Germany & University of Ljubljana, Slovenia)

This paper introduces the Corpus of Patient Documents-UK (CoPaDocs-UK; cf. Author 2025a), outlining its compilation, coverage, and potential for (socio)linguistic analyses of Late Modern English. CoPaDocs-UK is the sister corpus of the German CoPaDocs (`http://copadocs.de/`, cf. Schiegg 2022) and comprises letters written by individuals undergoing inpatient treatment in historical psychiatric hospitals in the second half of the nineteenth and the early twentieth centuries. These 'patient letters' were addressed to medical staff or recipients outside the hospitals (Schiegg 2020: 570), but they were subject to censorship: instead of being posted, many were filed away in casebooks, where they have survived to the present day.

For CoPaDocs-UK, patient records from six psychiatric hospitals in England and one in Scotland were searched for patient letters, both on-site in archives and, where digitised, online via the Wellcome Collection website (`https://wellcomecollection.org/`). A total of 639 patient letters were collected and transcribed diplomatically in XML with TEI specifications, retaining the original spelling and layout (e.g., page and line breaks, paragraphs, and indentation). The material dates from c.1840–1920, which reflects its general availability and data protection constraints. Extensive metadata on letter-writers, such as their year of birth, previous residence, occupation, education, religious affiliation, and diagnosis (if known), were extracted from case notes and compiled into a comprehensive database.

The 263 patients included in the corpus represent a broad socio-demographic spectrum of private, non-elite individuals, including, for example, governesses, accountants, and engineers as well as carpenters, miners, and general labourers – all born between the 1780s and 1890s. In the c.220,000 running words, there is a near gender balance with 44% written by women. In addition to patient letters, CoPaDocs-UK also contains around 140 further ego-documents either written by patients (e.g., notes, diary entries, and poems) or their relatives (mostly letters sent to the institutions).

CoPaDocs-UK offers a rare glimpse of authentic language use by private, non-elite individuals, complementing existing Late Modern English (correspondence) corpora, such as CONCE and CLMET. To illustrate the rich potential of CoPaDocs-UK for variationist historical sociolinguistics, the present paper briefly summarises my earlier studies on spelling variation and the decline of enregistered features of Quaker Plain Speech (e.g., THOU for standard YOU) – since one of the psychiatric hospitals from which the material was drawn was built for and run by Quakers (cf. Author 2025a, 2025b). Additionally, as a new line of research, the paper presents the potential of CoPaDocs-UK for morphosyntactic analyses through case studies on relativiser choice and the spread of not-contractions. By these case studies, I hope to demonstrate how CoPaDocs-UK broadens the empirical basis for research on variation and change in Late Modern English by looking beyond the language use of the educational elite and published authors.

References

Author. (2025a). English patient letters from the 19th and early 20th centuries: Variation and change in orthography and Quaker Plain Speech. Unpublished PhD thesis.

Author. (2025b). Plain speech and the Quaker pronoun of address in nineteenth-century England. In S. M. Litty & N. Langer (Eds.), Language ideology, policy, and practice: Focus on minoritized languages past and present (pp. 229–253). Lang.

Author. (In prep.). Apostrophes in Late Modern English nominal possessives: New evidence 'from below' based on patient letters.

CLMET = The Corpus of Late Modern English Texts, version 3.1, compiled by H. De Smet, S. Flach, H.-J. Diller, and J. Tyrkkö.

CONCE = A Corpus of Nineteenth-century English, compiled by M. Kytö and J. Rudanko.

CoPaDocs-UK = Corpus of Patient Documents: United Kingdom, compiled by Author and M. Schiegg, unpublished.

Schiegg, M. (2022). Flexible Schreiber in der Sprachgeschichte: Intraindividuelle Variation in Patientenbriefen (1850–1936). Winter.

RHINE

MOSELLE

Saturday
30 May 2026

Saturday 30 May 2026

Corpus-based Discourse Analyses E113 • 9:00–11:00

Degree Adverbs in Hong Kong English Newspaper Discourse FULL PAPER

Aditya Upadhyaya (University of Würzburg, Germany)

The interesting position of English in Hong Kong has sparked several studies on various features of Hong Kong English (HKE) (Peng and Setter, 2003; Joseph, 2004; Collins, 2009; Biewer et al. 2020). While some linguists suggest that the L2 speakers of English in Hong Kong still adopt their linguistic norms from British and/or American English (Luke and Richards 1982: 51-52; Hyland 1997: 206), others view it as a localized variety (Bolton 2000: 274-276; Sung 2015: 266-267). However, the limited availability of diachronic corpora of Hong Kong English has resulted in a predominance of synchronic studies of this variety, which fail to explain:
a) the differences between changes in genre development and varietal development (Noël and van der. Auwera, 2015) and,
b) "to what extent a new postcolonial variety of English has diverged across time from the historical input variety" (Mukherjee and Schilk 2012: 191).
The proposed study seeks to bridge these gaps by investigating the genre of newspaper writing in Hong Kong, using the diachronic corpus of Hong Kong English (DC-HKE). While news writing constitutes a realm of professional language use with articles discreetly edited, a more comprehensive investigation of newspaper discourse may uncover subtle changes unlikely to be edited out. This study is particularly concerned with how stance is portrayed in HKE newspaper discourse using degree adverbs, which express the assertion of the speaker and the degree of the word they modify. Stance-taking in HKE becomes an interesting premise, especially when considering the claims of Hong Kong media's self-censorship amidst political pressures and changing socio-political climate.
To investigate the extent to which HKE can be said to have diverged from its input variety, the BLOB, LOB and FLOB corpora of British English are used. The study seeks to investigate the following facets:
1. How has the frequency of degree adverbs in news discourse evolved over time? How far can HKE be said to have developed away from its input variety?
2. How do different factors like the type of degree adverb, journalist's role, and sentence level sentiment (adjectives) affect the use of degree adverbs in HKE newspaper discourse?
Based on Quirk et al.'s (1985) classification, 72 degree adverbs were extracted from DC-HKE

and British English corpora and coded for function and stance. Furthermore, sentiment analysis was performed using VADER (Valence Aware Dictionary and sEntiment Reasoner), which assigns polarity scores ranging from -1 (negative) to +1 (positive). Degree adverb–adjective pairs were scored as positive (> 0.05), negative (< -0.05), or neutral (between -0.05 and 0.05).

Preliminary results show that British English maintained a stable and higher use of degree adverbs over time, while HKE showed greater fluctuation and lower frequency. In HKE, especially in 1960s and 1990s, emphasizers and adjective amplification in negative contexts occur more often in reported speech, while neutral or positive contexts are more frequent in journalists' own stance. The ongoing research expects to uncover further changes in the genre and use of degree adverbs in HKE newspaper discourse due to changing socio-political climate in Hong Kong.

References

Biewer, C., Lehnen, L., & Schulz, N. (2020). "The future elected government should fully represent the interests of Hongkong people." Re-Assessing Modalising Expressions. Ed. Pascal Hohaus, and Rainer Schulze. Amsterdam: John Benjamins, 311–341.

Bolton, K. (2000). "The sociolinguistics of Hong Kong and the space for Hong Kong English." World Englishes 19.3: 265-285.

Collins, P. (2009). "Modals and quasi - modals in world Englishes." World Englishes 28.3: 281 -292.

Hyland, Ken. (1997). "Language attitudes at the handover: communication and identity in 1997 Hong Kong." English World-Wide 18.2: 191–210.

Joseph, J. E. (2004). "Case study 1: The new quasi-nation of Hong Kong." Language and identity: National, ethnic, religious 132-161. London: Palgrave Macmillan UK.

Luke, K. K., & Richards, J. C. (1982). "English in Hong Kong: functions and status." English World-Wide, 3.1: 47-64.

Mukherjee, J., & Schilk, M. (2012). "LOOKING INTO THE INTERNATIONAL CORPUS OF ENGLISH (ICE) AND BEYOND." The Oxford handbook of the history of English. Ed. Terttue. Nevalainen and Elizabeth Closs Traugott. New York: Oxford University Press. 189-194.

Noël, Dirk & van der Auwera, Johan (2015). "Recent quantitative changes in the use of modals and quasi-modals in the Hong Kong, British and American printed press." Grammatical Change in English World-Wide. Ed. Peter Collins. Amsterdam: John Benjamins. 437–464.

Peng, L., & Setter, J. (2000). "The emergence of systematicity in the English pronunciations of two Cantonese-speaking adults in Hong Kong." English World-Wide, 21.1: 81-108.

Sung, C. C. M. (2015). "Hong Kong English: linguistic and sociolinguistic perspectives." Language and Linguistics Compass 9.6: 256-270.

Large Language Models versus Varieties of English: A (very) short-term diachronic study

FULL PAPER

Julia Schlüter (University of Bamberg, Germany)

At a time when more and more English texts are influenced or even generated by AI tools (e.g. Liang et al. 2024, Kobak et al. 2025), an issue that corpus linguistics needs to address is to what extent this will have or already has repercussions on the evolution of the English language itself. Thus, the present paper pursues the following set of interrelated research questions: Where is the default output of Large Language Models (LLMs) situated with reference to geographical varieties of English, in particular the British and American standards? Is it possible to approximate authentic British (BrE) or American English (AmE) by appropriate prompting? How do different LLMs compare with regard to these questions? Are more recent versions of LLMs more or less successful than earlier ones at representing different target norms?

To answer these questions, three task types were constructed to simulate uses to which millions of users apply AI support on a daily basis. These include text correction, translation and cases of linguistic insecurity. Several prepositional expressions were chosen that exemplify one of the numerous areas in which British and American usage differ quantitatively (as attested in corpus studies by Algeo 2006: 159–198; e.g. *a lease of/on life*, *at/on short notice*, *membership of/in sth.*). Those that textbooks commonly draw attention to (e.g. *at/in school*, *at/on the weekend*, *in/on the street*) were deliberately excluded from the design. Three versions of each task were run, one without further specification, one with BrE and one with AmE set as targets, and the LLM output was monitored across 20 iterations to record probabilistic variation.

Results are compared between different LLMs (ChatGPT, Gemini, DeepSeek, Le Chat, Llama, Copilot, Claude) and between different versions of the LLMs (GPT4o-mini, GPT5; further updates on other models will follow as they become available in 2026). Usage data from the Corpus of Global Web-based English (GloWbE; Davies 2013) serve as a yardstick. The results indicate that, summed across all tasks, the tools show markedly stronger affinities with AmE (i.e. fewer corrections of AmE usage, more AmE-like productions) in the default case. Interestingly, the strength of the preference can be shown to depend on the task type. In addition, this tendency can be countered by explicit prompts defining BrE as a target. The size of this effect varies from one model to another, with DeepSeek and Gemini at present achieving the strongest differentiation between BrE and AmE, and Copilot and Le Chat scoring lowest. Furthermore, the current version ChatGPT5 shows no improvement in comparison to ChatGPT4o-mini, tested in early 2025.

The overarching interest of this research derives from concerns that massive use of LLMs will act as a homogenizing force and erode the diversity of (written and even spoken; Yakura

et al. 2024) Englishes, including standard BrE (Mair 2024). In conclusion, this scenario might become a reality, especially as many of the more fine-grained varietal differences remain below the level of individual or public attention.

References

Algeo, J. (2006). British or American English? A Handbook of Word and Grammar Patterns. Cambridge: Cambridge University Press.

Davies, M. (2013). Corpus of Global Web-Based English. Available online at https://www.english-corpora.org/glowbe/.

Kobak, D., González-Márquez, R., Horvat, E.-A. & Lause, J. (2025). Delving into ChatGPT usage in academic writing through excess vocabulary. Science Ad-vances 11(27), eadt3813. https://doi.org/10.1126/sciadv.adt3813.

Liang, W., Zhang, Y., Wu, Z., Lepp, H., Ji, W., Zhao, X., Cao, H., Liu, S., He, S., Huang, Z., Yang, D., Potts, C., Manning, C.D., & Zou, J.Y. (2024). Mapping the Increasing Use of LLMs in Scientific Papers. ArXiv, abs/2404.01268. https://doi.org/10.48550/arXiv.2404.01268.

Mair, Ch. (2024). ChatGPT for linguists: Source of data and copilot for analysis. Talk presented at the Hermann Paul Centre for Linguistics, University of Freiburg, 19 April 2024.

Yakura, H., Lopez-Lopez, E., Brinkmann, L., Serna, I., Gupta, P., & Rahwan, I. (2024). Empirical evidence of Large Language Model's influence on human spoken communication. ArXiv, abs/2409.01754. https://doi.org/10.48550/arXiv.2409.01754.

Will language creativity diminish in the age of AI-generated news content?

FULL PAPER

Andrew Kehoe, Matt Gee & Antoinette Renouf (Birmingham City University, United Kingdom)

This paper explores the effect AI-generated content is having on language variation in online news reporting. The use of Generative AI (GenAI) for content creation saw a huge surge with the release of ChatGPT in November 2022. Thirty years of research into linguistic change and neology has shown us that the language of news reporting is particularly creative. Now we address the question, as more and more online news content is generated by AI tools, will language use in this context 'stagnate'? Stagnation, in this sense, suggests that the lexicon will exhibit less change over time due to the proficiency of Large Language Models (LLMs) in replicating the content used in their training. Compounding this issue is the fact that LLMs will increasingly be trained on texts which have themselves been produced by previous generation LLMs, a phenomenon referred to as 'model collapse' (Shumailov et al. 2024). There is thus the risk of the lexicon being homogenised, with a reduction in lexical productivity and creativity. In this paper, we assess to what extent this phenomenon can already be observed.

Researchers have noted that LLM-generated texts in other domains often contain specific terms with higher than expected frequency (e.g. the use of 'delve', 'intricate', and 'underscore'

in academic articles: Juzek and Ward 2024). We have established in our own work a range of diachronic change patterns that words and neologisms may follow, and developed novel methods for tracking changes in the frequency and collocational environments of words over time (Authors 2012, 2022, 2024).

In this study, we draw on an existing 2 billion word diachronic corpus of UK news articles published prior to the proliferation of GenAI tools. We construct a comparison corpus of AI-generated articles via ChatGPT. To do so, we follow the two-stage process outlined by Juzek and Ward (2024), first creating summaries of human-authored articles from the existing corpus, and then from these generating articles of similar length to the original by AI means. We experiment with summaries of varying lengths and articles covering a diverse range of topics.

The analysis will identify the words over or under used in the AI-generated articles in comparison to human-authored texts. In addition, we compare lexical diversity (and other measures, e.g. Herbold et al. 2023) and the number of types at different frequency bands. We also discuss to what extent such differences can be observed over time with reference to the diachronic corpus as a whole.

We conclude by discussing the potential outlook of this research, the suitability of the methodology, and ways in which new corpora and methods may be designed to identify lexical stagnation.

References

Renouf, A. (2012). A Finer Definition of Neology in English: the life-cycle of a word. In Hasselgård, H., et al. (eds.). *Corpus Perspectives on Patterns of Lexis*. Amsterdam: John Benjamins.

Kehoe, A., Gee, M. & Renouf, A. (2022). A data-driven approach to finding significant changes in language use through time series analysis. In Flach, S. & Hilpert, M. (eds.) *Broadening the spectrum of corpus linguistics: New approaches to variability and change*. Amsterdam: John Benjamins.

Gee, M., Kehoe, A. & Renouf, A. (2024). Establishing a 'new normal': detecting fluctuating trends in word frequency over time. In Buschfeld, S., et al. (eds.) *Crossing Boundaries through Corpora: Innovative Approaches in Corpus-Linguistics*. Amsterdam: John Benjamins.

Herbold, S., Hautli-Janisz, A., Heuer, U., Kikteva, Z., & Trautsch, A. (2023). A large-scale comparison of human-written versus ChatGPT-generated essays. Scientific reports 13(1): 18617.

Juzek, T., & Ward, Z. (2024). Why Does ChatGPT "Delve" So Much? Exploring the Sources of Lexical Overrepresentation in Large Language Models. arXiv preprint arXiv:2412.11385.

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson R., & Gal Y. (2024). AI models collapse when trained on recursively generated data. Nature 631(8022): 755-759.

Evaluating the Performance of Large Language Models (LLMs) for Nomina Loci Extraction in Lithuanian

WIP

Danguole Straizyte (Vilnius University, Lithuania), Andrius Utka (Magnus University, Lithuania) & Martynas Sabaliauskas (Vilnius University,Lithuania)

The automated extraction of specific linguistic data is a fundamental aspect of contemporary corpus linguistics (McEnery & Hardie, 2012). However, in highly inflected languages such as Lithuanian, the automated identification of nomina loci presents a considerable challenge. Nomina loci are place nouns that denote derived names of places (e.g. dirbtuvė “workshop” derived from dirbti “to work”). The extensive morphological variation means that recognizing all instances of a specific place noun requires sophisticated handling of inflection and derivation. Manual extraction from large corpora is exceedingly labor-intensive.

While advanced Large Language Models (LLMs) offer promising results for many NLP downstream tasks (see e.g. named entity recognition (Šostaks et al., 2025)), their ability to handle the complexities of Lithuanian morphology remains to be systematically evaluated (Bergmanis, Pinnis, & Kapočiūtė-Dzikienė, 2025).

The present study investigates the capabilities and limitations of several popular LLMs for the task of automated collection of nomina loci from Lithuanian literary texts. We aim to address the following research questions:

1. What are precision, recall, and F1 scores of different state-of-the-art LLMs when extracting Lithuanian nomina loci from literary texts?
2. What are the main linguistic challenges, particularly concerning inflection and ambiguity, encountered during automated extraction, and how effectively do different models overcome them?
3. What methodological pipeline is most effective for utilizing LLMs for linguistic data collection in morphologically complex languages?

For the evaluation of LLMS, the research utilizes a ”gold standard” dataset comprising selected Lithuanian literary texts, which will be manually annotated by expert linguists, marking all instances of nomina loci.

Our approach involves a comparative analysis of the extraction results across different LLMs. The method relies on feeding the raw (unannotated) texts to each selected LLM using carefully engineered zero-shot or few-shot prompts designed for nomina loci identification. We will develop an automated evaluation pipeline, utilizing Python scripts to process the outputs from the models and to collect comprehensive statistics. The performance of each LLM will be measured by comparing its output against the ”gold standard” annotated dataset, allowing for the calculation of precision, recall, and F1 scores.

We anticipate that the results will demonstrate the varying efficacy of different LLMs for this specialized task. We hypothesize that the performance will differ significantly across models, highlighting the specific challenges posed by Lithuanian morphology even for advanced AI systems. Furthermore, a qualitative error analysis will categorize specific challenges, such as ambiguity between place nouns and common nouns, the correct handling of in-

flected forms, and the overall reliability of the AI-generated output. This study provides practical methodological insights for using advanced AI tools in corpus-based research on morphologically rich languages.

References

Kapočiūtė-Dzikienė, J., Bergmanis, T., Pinnis, M. (2025). Localizing AI: Evaluating Open-Weight Language Models for Languages of Baltic States. Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), 287-295.

McEnery, T., and Hardie, A. (2012). Corpus linguistics: Method, theory and practice. Cambridge University Press.

Šostaks, A., Rikačovs, S., Sprogis, A., Metra, O., and Lavrinovičs, U. (2025). Using LLM-s for Zero-Shot NER for Morphologically Rich Less-Resourced Languages. Baltic Journal of Modern Computing 13(2), 357-365.

Human-AI collaborative multimodal local grammar analysis of graphic data commentaries in medical journal articles

WIP

Ding Huang, Jiajin Xu & Yingming Song (Beijing Foreign Studies University, China, People's Republic of)

Graphic data commentary is a discourse act in academic writing that informs and discusses research data in visual forms such as tables, charts, graphs, and diagrams (Zhang et al. 2024). Previous research has examined graphic data commentary through design (e.g., Annesley 2010), rhetoric (e.g., information transfer in Widdowson [1979] and data commentary in Swales and Feak [2012]), and local grammar (Zhang et al. [2024] and Zhang and Zhang [2024]) perspectives. Notably, the fine-grained local grammar analysis demonstrates form-meaning patterning in graphic data commentary in economic discourse, but does not reveal correlations between graphic and data types or authors' choice in highlighting, reasoning, or elaborating specific data points in graphics. Therefore, the present study seeks to address three research questions:
(1) What is the local grammar of graphic data commentaries in medical journal articles?
(2) To what extent does the type of research data, a key functional element in graphic data commentary, differ across different graphic formats?
(3) In what ways do specific visual elements in graphics correspond to local grammar functional elements that describe or discuss graphic data elements?
We collected 250 tumour-related clinical research papers from four leading medical journals and randomly selected 50 for a pilot study. Sentences containing "fig" (based on referencing conventions observed in the four journals) were extracted from the main body of the articles, excluding figure captions. Graphics from both main texts and supplementary materials were included for comprehensive analysis.

To address the first research question, graphic data commentaries are annotated by both human researchers and a large language model (LLM). The human local grammar analysis uses functional labels adapted from Zhang et al. (2024). The LLM annotation proceeds in two parts: first, identifying the functional elements of RESEARCHER, ACT, HINGE, GRAPHIC, STUDY, ATTITUDE, and CONTENT in graphic data commentaries; second, distinguishing whether CONTENT represents DATA, INTERPRETATION, or ADDITIONAL INFORMATION based on the corresponding graphics. Human and LLM annotations are compared for consistency and reliability. To address the second question, we ask the LLM to classify the graphics (e.g., line charts) and determine the type of information they present (e.g., survival probability). The results are manually checked. To address the third question, we manually compared each commentary with its corresponding graphic to map DATA, INTERPRETATION, or ADDITIONAL INFORMATION elements to relevant visual elements.

The pilot study identified 11 local grammar patterns (≥ 5 instances respectively) in medical papers, which differed notably from those in economics papers in both form and distribution. Therefore, we expect similar findings in the full study, along with conventions linking graphic types to specific data presentation purposes. Moreover, mapping meanings expressed in graphic data commentaries to corresponding visual elements is expected to reveal synergetic cross-modal data presentation strategies in medical discourse.

This study contributes theoretically by describing verbal–visual mechanisms in medical research data presentation through fine-grained multimodal local grammar analysis. The study also contributes methodologically through the human-AI collaborative approach, and pedagogically by informing academic writing instruction to help medical students present and discuss research data more effectively.

References

Annesley, T. M. (2010). Put your best figure forward: line graphs and scattergrams. Clinical Chemistry, 56(8), 1229-1233.

Swales, J., & Feak, C. (2012). Academic Writing for Graduate Students: Essential Tasks and Skills (3rd ed.). Ann Arbor, MI: University of Michigan Press.

Zhang, L., Jiang, R., & Zhang, J. (2024). 'Table 1 shows that…': A local grammar of graphic data commentary in discourse of Economics. English for Specific Purposes, 74, 68–81. https://doi.org/10.1016/j.esp.2024.01.001

Zhang, L. & Zhang, Y. (2024). Tracking the changing patterns of graphic data commentary in economics research articles over time: A local grammar study. Journal of English for Specific Purposes, 72, 68–81. https://doi.org/10.1016/j.jeap.2024.101437

Widdowson, H. G. (1979). Explorations in Applied Linguistics. Oxford: Oxford University Press.

Fucking weird and bloody good: Recent changes in the use of taboo intensifiers in British English

FULL PAPER

Ulrike Stange-Hundsdörfer (JGU Mainz, Germany)

Taboo intensifiers are used to add extra force to an utterance (compare: a really good film vs. a fucking good film) and, like ordinary intensifiers (so good, really great) start losing this expressive force as soon as speakers use them (Bolinger 1972). Initially, new intensifiers combine with a limited number of lexical items, but as they gain in frequency, their original semantic meaning bleaches – a process known as delexicalization (Partington 1993, Lorenz 2002). Accordingly, taboo intensifiers like fucking or bloody first admit items with negative meaning only (fucking stupid, bloody awful) but in time become more flexible, combining also with neutral and even positive adjectives (fucking amazing, bloody brilliant; Aijmer 2018). Expressive intensifiers are transgressive in nature, breaking culture-specific taboos or norms (Farquaharson et al. 2020, Storch & Nassenstein 2020) and serve to express strong emotions or attitudes (Andersson & Trudgill 2007). Drawing their force from taboo-breaking, their patterns of use are also shaped by speaker demographics. Male speakers appear to have fewer inhibitions using taboo language than female speakers (Jay 1992, Mehl & Pennebaker 2003), and expressive intensifiers appear characteristic of younger people (Palacios Martínez & Núñez Pertejo 2012, Aijmer 2018).

The present study replicates Aijmer's corpus study on fucking (Aijmer 2018) and addresses this main research question: Who uses bloody and fucking as adjective intensifiers and what changes can be detected in their use in spoken British English in the past 30 years? Aspects covered include speaker groups, frequencies, collocational patterns and syntactic distribution. Additionally, survey data will shed light on potential changes in attitude towards these two taboo intensifiers.

The corpus analyses draw on data from the demographically sampled section of BNC1994 (5m words, Nbloody=540, Nfucking=337) and the Spoken BNC2014 (11.4m words, Nbloody=337, Nfucking=790). First analyses reveal that frequency-related findings from Aijmer's study (2018) on fucking could not be replicated, resulting in a different story about relevant changes: it is most frequently used by speakers aged 25-34 (normalised frequencies), regardless of gender (gender effect has disappeared, log ratio=-0.25). Bloody is decreasing in frequency across speaker groups (from 108 to 19 occ. pmw), while fucking, stable overall (67 and 69 occ. pmw), is decreasing for male speakers (from 127 to 65 occ. pmw) but increasing for female speakers (from 21 to 70 occ. pmw). This suggests that bloody is on its way to the intensifier recycling bin (cf. Ito & Tagliamonte 2003), and that female speakers have caught up with male speakers regarding the use of fucking. Accordingly, bloody is attested with fewer types than in earlier data (from c. 225 to c. 120), while fucking continues expanding its collocational range (from c. 160 to c. 270). Thus, in contradiction to Aijmer's (2018)

findings, bloody is not a "close competitor" to fucking, nor are they the "most frequent intensifier[s] after very, really and so". Also, the disappearance of the gender effect for fucking is indicative of blunting (and/or emancipation or bonding among females, cf. Aijmer 2018: 74) at work. Survey data (to be analysed) will clarify this.

References

Aijmer, K. (2018) That's well bad. Some new intensifiers in spoken British English. In: V. Brezina, R. Love & K. Aijmer (eds.): Corpus Approaches to Contemporary British English, 60-95. New York: Routledge.

Andersson, L. & P. Trudgill (2007) Swearing. In: Monaghan, L.F.; Goodman, J.E. & J.M. Robinson (eds.) A Cultural Approach to Interpersonal Communication: Essential Readings, 195–199. Oxford: Blackwell.

Bolinger, D. (1972) Degree words. The Hague: Mouton.

Farquharson, Joseph T; Forrester, Clive and Andrea Hollington (2020) The Linguistics of Jamaican Swearing: Forms, Background and Adaptations. In Nassenstein, N. & A. Storch (eds.) Swearing and Cursing: Contexts and Practices in a Critical Linguistic Perspective, 147–164. Berlin/Boston: De Gruyter Mouton.

Ito, R. & S. Tagliamonte (2003): Well weird, right dodgy, very strange, really cool: Layering and recycling in English intensifiers. In: Language in Society 32: 257-279.

Jay, T. (1992) Cursing in America: A Psycholinguistic Study of Dirty Language in the Courts, in the Movies, in the Schoolyards and on the Streets. Philadelphia: John Benjamins Publishing.

Lorenz, G (2002) "Really worthwhile or not really significant?" A corpus-based approach to the delexicalisation and grammaticalization of intensifiers in Modern English. In Wischer, I. & G. Diewald (eds): New reflections on grammaticalization, 143-161. Amsterdam: John Benjamins.

Palacios Martínez, I. & P. Núñez Pertejo (2012) He's absolutely massive. It's a super day. Madonna, she's a wicked singer. Youth language and intensification: A corpus-based study. In: Text and Talk 32 (6): 773-796.

McEnery, A. & Z. Xiao (2004) Swearing in modern British English: the case of fuck in the BNC. Language and Literature 13: 235.

Mehl, M.R. & J.W. Pennebaker (2003) The Sounds of Social Life: A Psychometric Analysis of Students' Daily Social Environments and Natural Conversations. Journal of personality and social psychology 84 (4): 857–870.

Partington, A. (1993) Corpus evidence of language change. The case of the intensifier. In Baker, M., G. Francis & E. Tognini-Bonelli (eds.): Text and technology: In honour of John Sinclair, 177-192.Amsterdam: John Benjamins.

Storch, A. & N. Nassenstein (2020) 'I will kill you today' - Reading 'bad language' and Swearing through Otherness, Mimesis, Abjection and Camp. In Nassenstein, N. & A. Storch (eds.) Swearing and Cursing: Contexts and Practices in a Critical Linguistic Perspective, 1–36. Berlin/Boston: De Gruyter Mouton.

Recent diachronic change in affiliative vocatives in British English: A corpus-based study

FULL PAPER

Mariam Gagua (University of Bonn, Germany), Lisa Altendorf (University of Bonn, Germany), Christina Nelson (University of Bonn, Germany), Philipp Meer (University of Münster, Germany) & Robert Fuchs (University of Bonn, Germany)

Affiliative vocatives – a cover term for forms like dude, girl, or mate ('familiarizers'), darling, love, or honey ('endearments'), and bro or sis ('kinship-derived terms') – are often used not only to attract or maintain attention, but also to signal stance or manage interpersonal relationships (Leech, 1999). Affiliative vocatives are underexplored in British English, where research has largely been qualitative or focused on individual vocative forms (Baumgarten 2021, 2023; Palacios Martínez 2021; Pastorino 2022). Insights into large-scale patterns of variation and change, as reported in other varieties of English (Needle & Tagliamonte 2025), are currently not available.

To address this gap, we investigate short-term diachronic change in affiliative vocative use in British English (BrE) from the 1980s/90s to the 2010s by comparing data from the original British National Corpus (BNC1994) with the recent BNC2014 (Love et al. 2017). Specifically, we focus on the following research questions:

RQ1 What is the distribution of affiliative vocatives in contemporary BrE?

RQ2 What short-term diachronic changes (if any) have occurred in BrE?

RQ3 How does the use of affiliative vocatives in BrE vary by macro-sociolinguistic variables?

Based on a comprehensive list of vocatives taken from Needle & Tagliamonte (2025), we subjected all potential items to a Subject Matter Expert rating procedure, in which four linguists independently identified affiliative vocatives based on theoretically motivated criteria of social closeness, informality, equality, and non-hierarchical address, excluding items that were flirtatious, parental, or deferential in tone. This process yielded a final list of 32 items.

We extracted concordance lines for al 32 items in LancsBox (Brezina et al. 2021), along with the associated speaker and text metadata. Each occurrence was manually inspected to determine whether it functioned as a vocative; tokens used referentially, in reported speech, in formulaic material, or as affective interjections were discarded. Interrater reliability showed substantial agreement in both corpora ($\kappa = 0.777$ for BNC1994, $\kappa = 0.642$ for BNC2014). The final dataset comprised 6,314 affiliative vocative tokens (2,748 in BNC1994; 3,566 in BNC2014). Statistical modeling made use of Boruta random forests and generalized linear mixed-effects modeling.

Results reveal an increase in masculine-derived familiarizers and a concurrent decline in traditional endearments (see Figures 1-3). Younger speakers lead the shift toward newer familiarizers (mate, bruv, bro, dude) that diffuse across genders (see Figure 4), while women continue to favor endearments overall. These findings suggest a functional reorganization of affiliative address in contemporary British English, reflecting broader societal trends toward informality, partial gender convergence, and the redefinition of interpersonal warmth in

everyday interaction.

References

Baumgarten, N. (2021). Love as a term of address in British English: Micro-diachronic variation. Contrastive Pragmatics, 3(1), 31–58. https://doi.org/10.1163/26660393-BJA10027

Baumgarten, N. (2023). And the postcode darlin': Vocative variation in service encounters on the telephone in Northern England. In N. Baumgarten & R. Vismans (Eds.), It's different with you: Contrastive perspectives on address research (pp. 220–244). Amsterdam: John Benjamins.

Brezina, V., Weill-Tessier, P., & McEnery, T. (2020). LancsBox v. 6.x [Computer software]. Lancaster University. https://corpora.lancs.ac.uk/lancsbox

Leech, G.N. (1999). The distribution and function of vocatives in American and British English conversation. In H.Hasselgård & S.Oksefjell (Eds.), Out ofCorpora:Studies in Honour of Stig Johansson (pp.107–118). Amsterdam: Rodopi.

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken British National Corpus 2014: Designing and building a spoken corpus of everyday conversations. International Journal of Corpus Linguistics, 22(3), 319–344. https://doi.org/10.1075/ijcl.22.3.02lov

Needle, J. M., & Tagliamonte, S. A. (2025). Buddies, dudes, and bros of Ontario: Trends and patterns of vocative change. Journal of English Linguistics, 53(2), 107–136. https://doi.org/10.1177/00754242251341318

Palacios Martínez, I. M. (2021). Taboo vocatives in the language of London teenagers. Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA), 31(2), 250–277. https://doi.org/10.1075/prag.19028.pal

Pastorino, V. (2022). Dude in British English: Towards a non-gendered term of address. York Papers in Linguistics, 2(17), 13–28.

The "Spirit" in the Chamber: Capturing Zeitgeist in UK parliamentary debates

WIP

Penelope Gia Bao Huu Nguyen (University of Sheffield, United Kingdom)

Zeitgeist, or the "spirit of the time," is a concept commonly used in historical and sociological research. However, its definition remains vague, domain-dependent, and usually lacks bottom-up empirical evidence, especially from large-scale datasets. This lack of empirical grounding is a significant issue, as humanities scholars are often criticized for biases and a "cherry-picking" of evidence. Furthermore, the Zeitgeist, as a significant yet abstract force, is likely to both influence and be reflected by important semantic phenomena over time. In this study, Zeitgeist is redefined as the cultural, political, moral, and intellectual climate of a time period, realized by a certain group of language users by means of discourse. The Zeitgeist, in this case, will be determined through proxies, which are lists of keywords and key semantic domains. This mixed-methods study draws upon recent advancements in corpus linguistics. The process begins with the determination of keywords and key semantic categories using a keyness analysis. Following Gries (2024), this keyness will be treated as

a weighted, three-dimensional measure rather than a monodimensional one, using the versatile Kullback-Leibler Divergence (KLD) (Kullback & Leibler, 1951) as the core statistical measure. Additionally, as a diachronic study, the data from the previous period is treated as the reference corpus of the past, whereas that of the period right after is the target corpus of the new time. A key innovation of this project is the incorporation of semantic tags into the pipeline, providing another dimension of information. Further scrutiny of the concordance lines, enhanced by encyclopedic knowledge, is conducted if necessary to shed light on areas that the quantitative approach fails to explicate. The expected outcome is a list of keywords with key semantic categories for each time slice, and together, they co-construct the changing Zeitgeist in British parliamentary debates.

The data comes from the Semantic Hansard corpus, the largest publicly available corpus which contains nearly all speeches given in the UK Parliament spanning from 1803 to 2005 and is annotated by the USAS tagger (Rayson, 2008) with semantic tags. This unique source is particularly suited for capturing the Zeitgeist as it represents the significant concerns within a national legislative body, which in turn aims to represent broader public concerns. In contrast, more specialized or domain-dependent datasets often lack the longitudinal scale and broad representative scope required to provide the bottom-up empirical grounding necessary for a macroscopic cultural analysis. The preliminary analysis focuses on debates occurring during Thatcher's, Callaghan's, and Blair's governments.

The study's contributions are theoretical, methodological, and empirical. Theoretically, it refines the abstract concept of Zeitgeist into a data-backed, empirically verifiable construct, informing both conceptual history and corpus linguistics. Methodologically, the study evaluates the KLD as a versatile, bottom-up statistical tool, especially when used in conjunction with rich semantic annotation. Empirically, this research will deliver concrete, data-driven accounts of macroscopic British political Zeitgeists across governments.

References

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

Kullback, Solomon & Richard A. Leibler. 1951. On information and sufficiency. The Annals of Mathematical Statistics 22(1). 79–86.

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

A diachronic corpus-assisted study of the use of the term overtourism in UK national newspapers

WIP

Erik Castello (University of Padua, Italy) & Dario Del Fante (University of Ferrara, Italy)

The term "overtourism" has recently become a buzzword to talk about a phenomenon that is not new in tourism scholarship. It refers to tourism-led overcrowding in areas previously

used exclusively by residents and to how such situations are perceived by various stakeholders with different agendas (Yrigoya et al. 2024). The Oxford English Dictionary does not yet include an entry devoted to it, which suggests that the exact definition of its denotation and connotation still needs scholarly attention. This study sets out to explore the linguistic construction of overtourism in the British news media, with a special focus on its use over time and on the grammar of its representation in news discourse (van Leeuwen, 2008). Ultimately, it aims to inform the creation of corpus-assisted activities to engage language students in linguistic explorations of patterns of use of specific terms and in conversations about tourism (and other social phenomena) and their impact on the environment (Poole 2024).

It answers three main research questions: How is the term overtourism used diachronically in UK newspapers? Who are the main social participants (actors, places and concepts) revolving around it and how are they represented linguistically? What social actions do they perform and what roles are they given? The study adopts corpus linguistic and CADS methods (Baker 2023, Partington et al. 2013), with a view to identifying keywords and lexico-grammatical patterns that encode representations of the phenomenon and of related participants and processes. It is based on a corpus of about 3,100,000 tokens, specially compiled from UK National newspapers through Nexis Uni, using the search terms overtourism/over-tourism. The software used is #LancsBox X 5.0.0 (Brezina and Platt 2024).

A preliminary analysis shows that its first attested uses date to May 2017. Its presence increased until 2019, decreased until 2022 due to the pandemic, and has been growing exponentially since 2024. Overtourism is the most used orthographic form, yet over(-)tourism is also employed, especially in the early years. A diachronic analysis of collocates suggests that in 2017 the meaning of the term is often explained, in 2018-2019 the impacts of activities such as cruising are discussed, while lately journalists often report on protests against it and on proposals to tackle its negative effects. It also reveals the presence of likely synonyms of the term (e.g. anti-tourism, unsustainable tourism) and antonyms (e.g. responsible/sustainable tourism). An inspection of a selection of concordance lines indicates that the main social agents are impersonal groups (e.g. locals, protesters), businesses (e.g. hoteliers, Airbnb), and names of places that are used metaphorically (e.g. destinations, Venice). These tend to take active roles (e.g. march against, complain about) and less frequently passive ones (e.g. affected by, suffer from).

The work-in-progress report will discuss these and other findings, offering corpus-driven insights into how news media shape the perception(s) of overtourism.

References

Baker, P. (2023). Using Corpora in Discourse Analysis, London: Bloomsbury.

Brezina, V., Platt, W. (2024). #LancsBox X 5.0.0 [software package], lancsbox.lancaster.ac.uk

Partington, A., Duguid, A., & Taylor, C. (2013). Patterns and Meanings in Discourse: Theory and Practice in Corpus-Assisted Discourse Studies (CADS), Amsterdam & Philadelphia: John Benjamins.

Poole, R. (2024). A corpus-assisted ecolinguistic analysis of hurricanes and wildfires and the

potential for corpus-assisted eco-pedagogy in ELT classrooms. In M. Bortoluzzi & E. Zurru (Eds.), Environment as Lifescape: Ecoliteracy and Communication in Action. Bloomsbury.

van Leeuwen, T. (2008). Discourse and Practice: New Tools for Critical Discourse Analysis, Oxford: Oxford University Press.

Yrigoy, I., Horrach, P., Escudero, L., & Mulet, C. (2024). Co-opting overtourism: tourism stakeholders' use of the perceptions of overtourism in their power struggles, Journal of Sustainable Tourism, 32:4, 818-834,

Lexical and Ideological Change in Manifestos: A Diachronic Lexical and Ideological Change Analysis of Key Political Term Collocations in UK General Election Manifestos (1979-2024)

WIP

Mert Yeşilyurt (Ludwig-Maximilians-Universität München, Germany)

This study investigates whether diachronic lexical and ideological change can be detected through the collocations of key political terms in UK general election manifestos (1979–2024). It aims to address three research questions:
1) Do the collocational profiles of politically key terms such as freedom, policy, and tax reveal lexical and ideological change?
2) Do shifts in collocational profiles correspond to ideological trends across distinct political eras?
3) Do election manifestos reveal identifiable patterns of lexical innovation and diffusion?
The dataset comprises manifestos for six parties, Conservative, Labour, Liberal/Liberal Democrat, Plaid Cymru, Scottish National Party, and Sinn Féin, assembled in Sketch Engine and supplemented from ManifestoVault. To highlight lexical and ideological change, the corpus is partitioned into three era subcorpora chosen to capture major political events. Era 1 (1979-1992) corresponds to Thatcherism, while Era 2 (1997-2010) highlights the New Labour and the Coalition period, and Era 3 (2015-2024) highlights the Brexit and post-Brexit period. This partitioning emphasizes the periods' dominant ideological reorientations.
Methodologically, the study combines quantitative collocation metrics with qualitative concordance analysis. Texts are POS-filtered to focus on noun collocates, and extraction uses a two-word left-context window (–2) to capture fixed expressions and tightly bound collocations. The study measures association strength with logDice, indexes prevalence as frequency per million tokens, and employs log-likelihood as a simple significance check. For each target lemma, the five strongest collocates per era and per party are identified and then subjected to concordance inspection and historical contextualization, so that quantitative signals are interpreted against political events and discourse histories.
The results are expected to indicate that collocational patterns reliably track both lexicalization and ideological reorientation. Quantitatively robust collocates (high frequency per million, LL, and logDice) supported by concordance evidence show freedom moving from an individual-liberty frame in Era 1 through market-oriented framings in Era 2 to

sovereignty and post-Brexit framings in Era 3. Policy collocates shift from defense and security emphasis toward sectoral concerns, notably fishery, and then toward broader foreign policy orientations. Tax displays both thematic continuity and lexical innovation through collocates such as “council tax”, “stealth tax”, and “bedroom tax” emerge, spread, and become conventionalized across manifestos.
Overall, the collocation metrics with concordance-based close reading demonstrates that manifestos are a tractable and sensitive resource for tracing diachronic lexical and ideological change in political discourse.

References

Atkinson, A. B. (2004). Income Tax and Top Incomes over the Twentieth Century. Hacienda Pública Española / Review of Public Economics, 168(1), 123–141.

Diamond, P., Newman, J., Richards, D., Sanders, A., & Westwood, A. (2024). ‘Hyper-active incrementalism’ and the Westminster system of governance: Why spatial policy has failed over time. The British Journal of Politics and International Relations, 26(4), 1185–1210. https://doi.org/10.1177/13691481241259385

Dolinsky, A. O., Huber, L. M., & Horne, W. (2025). ManifestoVault V1.0: Annotated full-text general election manifestos at the natural-sentence level of three European countries 1970 2025 (Version 5) [Dataset]. DataverseNL. https://doi.org/10.34894/VKQSPO

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. Comput. Linguist., 19(1), 61–74.

Evert, S. (2004). The Statistics of Word Cooccurrences: Word Pairs and Collocations (Dissertation).

Fairclough, N. (with NetLibrary, Inc). (2003). Analysing discourse: Textual analysis for social research. Routledge.

Fuchs, C. (2016). Neoliberalism in Britain: From Thatcherism to Cameronism. tripleC: Communication, Capitalism & Critique. Open Access Journal for a Global Sustainable Information Society, 14(1). https://doi.org/10.31269/triplec.v14i1.750

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

Hills, J. R. (2002). The welfare state in the UK: Evolution, funding and reform. https://www.ipss.go.jp/webj-ad/webjournal.files/socialsecurity/2002/02mar/JohnHills.pdf

Jessop, B. (2004). From Thatcherism to New Labour: Neo-Liberalism, Workfarism, and Labour Market Regulation.

Kalaš, F. (2025). Quantifying Lexical Shifts in Political Speech: A Corpus- Based And AI-Driven Analysis of Power and Influence.

McEnery, T., & Hardie, A. (2012). Corpus linguistics: Method, theory and practice. Cambridge University Press.

Pearce, M. (2004). The marketization of discourse about education in UK general election manifestos. Text - Interdisciplinary Journal for the Study of Discourse, 24(2). https://doi.org/10.1515/text.2004.009

Pearce, M. (2014). Key Function Words in a Corpus of UK Election Manifestos. Linguistik Online, 65(3). https://doi.org/10.13092/lo.65.1402

Phillipson, J., & Symes, D. (2018). "A sea of troubles": Brexit and the fisheries question. Marine Policy, 90, 168–173. https://doi.org/10.1016/j.marpol.2017.12.016

Ramsbotham, O., & Miall, H. (1991). The British Defence Debate in the 1980s. In O. Ramsbotham & H. Miall, Beyond Deterrence (pp. 127–143). Palgrave Macmillan UK. https://doi.org/10.1007/978-1-349-21720-5_6

Rychlý, P. (2008). A Lexicographer-Friendly Association Score. RASLAN.

Schelkle, W., Kyriazi, A., Ganderson, J., & Altiparmakis, A. (2024). Brexit – the EU membership crisis that wasn't? West European Politics, 47(5), 997–1020. https://doi.org/10.1080/01402382.2024.2325780

Schmid, H. & Mantlik, A. (2015). Entrenchment in Historical Corpora? Reconstructing Dead Authors' Minds from their Usage Profiles. Anglia, 133(4), 583-623. https://doi.org/10.1515/ang-2015-0056

Schmid, Hans-Jörg, The Dynamics of the Linguistic System: Usage, Conventionalization, and Entrenchment (Oxford, 2020; online edn, Oxford Academic, 20 Feb. 2020), https://doi.org/10.1093/oso/9780198814771.001.0001

Traugott, E. C. (2017). Semantic Change. In E. C. Traugott, Oxford Research Encyclopedia of Linguistics. Oxford University Press. https://doi.org/10.1093/acrefore/9780199384655.013.323

Yesilyurt, M. (2025). party_manifestos [Corpus created in Sketch Engine]. Retrieved July 28, 2025, from https://ske.li/ukmanifestos

Corpus Grammar Research [2] E314 • 9:00–11:00

Quantifying the Productivity of Present-day English Derivational Suffixes: A corpus-based case for Shannon's entropy as a measurement of productivity

FULL PAPER

Tamás Fekete (University of Pécs, Hungary)

In different treatments of morphological productivity it is assumed that type frequency, especially the frequency of hapax legomena, can be a reliable indicator of how productive a given word-formation method is (e.g. Baayen & Lieber 1991). In itself, however, hapax count is not sufficient for determining the true productivity of a word formation rule. The goals of this paper are twofold: on the one hand it aims to demonstrate the usefulness of Shannon's entropy (Shannon 1948) as a measure of productivity and on the other to provide a corpus-based analysis of the productivity of present-day English derivational suffixes with the help of entropy. Thus, the research questions this paper aims to answer are: (i) how can entropy incorporate existing measurements of morphological productivity and (ii) what differences are observable in the productivities of present-day English derivational suffixes? Baayen (2009) introduces three different measurements of productivity: (i) realized productivity, which is essentially equal to the sheer number of words in a given affix, (ii) expanding

productivity, which aims to quantify at what rate can a word-formation process produce new members by dividing the hapax count of a given affix by the total hapax count in the corpus, and (iii) potential productivity which measures to what extent a morphological category is saturated by calculating the hapax-to-token ratio of a given affix. In the present paper, it is argued that entropy is a useful tool for the measurement of morphological productivity, as it incorporates all three of Baayen's metrics while also counterbalancing their caveats.

The research was carried out on the 4-million-word Baby version of the British National Corpus. Altogether 53 suffix types were selected for analysis on the basis of Marchand (1969), and 282,234 suffixed word tokens were collected from the 4-million-word BNC Baby, which correspond to a total of 12,888 suffixed word types. As productivity is a continuum of word-formation rules ranging from fully transparent processes on one extreme to fully lexicalized and opaque ones on the other, words in each suffix were split into two categories: words with free bases and words with obligatorily bound or lexicalized base, yielding a total of 123 suffix tokens, pertaining to the 53 suffix types. Entropy was then calculated for each suffix token and the ratio between the entropy values of bound-base suffix tokens and free-base suffix tokens within each suffix type was computed. This ratio was then compared across the different suffix types, resulting in the final productivity scale. It is assumed that the higher the bound-to-free base entropy ratio is, the more lexicalized tokens are found in the given suffix, which corresponds to a lower degree of productivity, while a lower bound-to-free ratio implies higher productivity.

In general, verbalizers emerge as the least productive category of suffixes with an overall bound-to-free ratio of 1.15, meaning that lexicalized words outnumber non-lexicalized items. In the case of adverb-forming suffixes, deadjectival -ly dominates the entire category and occurs solely with free bases, while nominalizers come in at a ratio of 0.8 and adjectivalizers at 0.7.

References

Baayen, H. (2009). Corpus linguistics in morphology: Morphological productivity. In A. Lüdeling & M. Kytö (Eds.), Corpus linguistics: An international handbook, Volume 2. (pp. 899–919). De Gruyter.

Baayen, H., & Lieber, R. (1991). Productivity and English derivation: a corpus-based study. Linguistics, 29(5), 801–843.

BNC Consortium. (2005). The BNC Baby. http://www.natcorp.ox.ac.uk/

Marchand, H. (1969). The categories and types of present-day English word-formation. Second edition. C. H. Beck.

Shannon, C. E. (1948). A mathematical theory of communication. The Bell System Technical Journal, 27, 379–423, 626–656.

Squandering the taxpayers money? Variation in English genitive usage

FULL PAPER

Patricia Ronan (TU Dortmund, Germany), Gerold Schneider (University of Zurich) & Martin Schweinberger (University of Queensland)

Genitive variation such as New Year's message versus message of the New Year and factors influencing it are a much discussed topics in World Englishes and in corpus linguistics in particular and it has been found that that genitive variation is impacted by various features such as register and variety of English (e.g. Biber et al. 2023, Szmrecsanyi, Grafmiller, Heller & Röthlisberger 2016). It has furthermore been noted that not only of genitive and the Anglo-Saxon 's genitive need to be considered in a the study of the envelope of variation, but also noun-noun modification like the New Year message (Smrecsanyi et al. 2016) form part of the envelope of variation. In a study of Late Modern English covering the period until 2000, Smrecsanyi, Biber, Egbert & Franco (2016) find that the latter is on the rise compared to 's genitive and of-genitive.

Work on the distribution of noun-noun modification like New Year message (e.g. Rosenbach 2019), does not focus on the variation between plural noun – headnoun and plural Saxon genitive – headnoun, such as students' union versus students union, citizens' advice versus citizens advice or taxpayers' money versus taxpayers money, even though their homophony makes this a particularly interesting field of study.

The current study helps to bridge this gap. It pursues the research question which varieties of English favour the use of premodifying plural nouns over premodifying plural genitive nouns? To answer this question, it uses the offline-version of the GloWbE corpus (Davies 2013) and the tag sets [nn2] ' [nn] are compared to [nn2] [nn] to study the distribution of these structures in the geographic varieties represented in GloWbE. The focus of our study is on L1 varieties of English to minimize influence of language contact on the results.

Results are expected to show that Australian English, together with Irish and British English, is leading in the use of noun (gen.pl) – noun compounds, while American English, amongst others, continues to favour Anglo-Saxon genitives. However, the distribution varies with individual lexical items, which may signal frequency effects.

References

Biber, Douglas, Benedikt Szmrecsanyi, Randi Reppen & Tove Larsson. 2023. Expanding the scope of grammatical variation: towards a comprehensive account of genitive variation across registers. English Language and Linguistics 28(1):1-39.

Davies, Mark. 2013. Corpus of Global Web-Based English. Available online at `https://www.english-corpora.org/glowbe/`.

Rosenbach, Anette. 2019. On the (non-)equivalence of constructions with determiner genitives and noun modifiers in English. English Language and Linguistics, 23.4: 759–796.

Szmrecsanyi, Benedikt, Douglas Biber, Jesse Egbert & Karlien Franco. 2016. Toward more accountability: Modeling ternary genitive variation in Late Modern English. Language Variation

and Change 28(01):1-29.

Szmrecsanyi, Benedikt, Jason Grafmiller, Benedikt Heller & Melanie Rothlisberger. 2016. Around the world in three alternations: Modeling syntactic variation in varieties of English. English World-Wide 37(2), 109–37.

The Nuts and Bolts of English Collective-Verb Agreement: Studying the Interplay of Semantics, Syntax and Complexity in Concord

FULL PAPER

David Hernández-Coalla (Universidade de Vigo, Spain)

The capacity of English collective nouns to prompt both singular and plural concord with their agreeing verbs is mentioned in all main reference grammars (Quirk et al. 1985: 757-759, Biber et al. 1999: 190-191, Huddleston & Pullum 2002: 501-504). However, the explanations provided about the nature of this phenomenon have traditionally revolved around a simple contrast between unity-profiling and distributive readings. Recent research has questioned this assumption in light of the potential role of a wide variety of factors, such as the distance between the collective and the target verb (Levin 2001: 92-99), the particular characteristics of collectives partaking in binomial phrases (Fernández-Pena 2020) or the semantic type of the verb (Lakaw 2024: 166-172). However, to date there is no detailed quantitative account of concord with bare collectives (cf. "a family" and "a bunch of NOUNPL") that considers their recent evolution alongside all classes of predictors: semantic, syntactic, complexity-related, extra-linguistic, etc. This study proposes a corpus-driven analysis of the agreement patterns of eight collective nouns with significant rates of variable agreement ('audience', 'band', 'cast', 'club', 'crowd', 'family', 'gang' and 'team') based on contemporary data from British English, the variety where plural agreement is said to be most frequent. My research questions are:

- Do certain factors have an impact on the number of the agreeing verb?
- To which extent are the agreement patterns of collective-verb pairs influenced by register?

To this end, data were retrieved from the British National Corpus (BNC Consortium 2007) and the British National Corpus 2014 (Love et al. 2017, Brezina et al. 2021), since their makeup allows for the comparison of results both for the diachronic and register-based components of the analysis. 500-instance samples were extracted for each collective noun and corpus whenever possible, for a total of 16 samples and almost 6,000 tokens. These were all annotated for 13 predictors that cover the whole range of variable types that may condition agreement. The statistical analysis, carried out using R (R Core Team 2024), is intended to endow the results with a solid quantitative account based on multivariate models: linear regression (checked for the possible collinearity of predictors) and random forests.

Preliminary research with a reduced set of samples points to a slight increase in plural agreement over the last decades and the relative importance of the intervening linguistic material between the noun and the verb, the semantic category of the latter and the type of determiner. Substantial effects were also attested for the register variable, with a marked presence of plural agreement in oral discourse compared to more structured written texts.

References

BNC Consortium. (2007) The British National Corpus, XML edition. Oxford Text Archive. http://hdl.handle.net/20.500.14106/2554.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan E. (1999). Longman grammar of spoken and written English. Longman.

Brezina, V., Hawtin, A., & McEnery, T. (2021). The written British National Corpus 2014 – design and comparability. Text & Talk, 41 (5-6), 595-615.

Fernández-Pena, Y. (2020). Reconciling synchrony, diachrony and usage in verb number agreement with complex collective subjects. Routledge.

Huddleston, R., & Pullum, G. K. (2002). The Cambridge grammar of the English language. Cambridge University Press.

Lakaw, A. (2024). Agreement with collective nouns: Diachronic corpus studies of American and British English. PhD dissertation. Linnaeus University Press.

Levin, M. (2001). Agreement with collective nouns in English. Department of English of Lund University.

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.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). A comprehensive grammar of the English language. Longman.

R Core Team. (2024). R: A language and environment for statistical computing. https://www.r-project.org/

Clearly and obviously in spoken interaction

FULL PAPER

Karin Aijmer (University of Gothenburg, Sweden)

Clearly and obviously are evidential adverbs which are sometimes difficult to distinguish from each other. Clearly has been discussed much less than obviously which is to be expected since it is less frequent as an evidential adverb than obviously. Specifically, it belongs to a category of evidential markers derived from manner adverbs (cf. Carretero 2019). In (1) clearly is a manner adverb with the meaning of 'in a clear way'.
(1)Do you wanna speak clearly please?
The evidential meaning is illustrated in (2):
(2) well clearly the authorities are really worried about a terrorist attack

In the evidential use clearly does not express perception directly but via inference ('it is clear that'). It is distinguished from the manner adverb by its position in the clause and propositional scope.
Previous studies of clearly have analysed the adverb in writing. Alonso-Almeida (2012) investigates the functions of clearly and obviously in research articles. Rozumko's corpus-based study (2019) focuses on the functions of clearly and obviously with a second person pronoun subject in rhetorical texts.
The present article investigates the forms and functions of clearly in the Spoken BNC 2014 (Love et al. 2017) and makes comparisons with obviously in that corpus. Clearly as an evidential adverb is attested already in the 16th century according to the Oxford English Dictionary. It is increasing in frequency over a short time in present-day English. In comparable samples of five million words from the British National Corpus there were 12 cases of the evidential clearly to be compared with 173 cases of the adverb in the Spoken BNC2014.
The evidential adverbs need to be analysed from a broad interactive perspective involving pragmatic issues such as the reference to certainty and who has the right to knowl. For example, speakers use clearly to express their attitudes to an event or object talked about and they can use the adverb to upgrade a claim.
The findings of the corpus investigation suggest that we can distinguish between cases where clearly has inferential meaning associated with certainty and examples where it is used by the speaker to take up an epistemic stance showing the hearer how certain or confident she (cf. Dendale 2020). The meaning of clearly is close to definitely and has the pragmatic effect that the claim to truth is upgraded or emphasized (something is the only alternative). In the Spoken BNC2104 clearly often competes with obviously in the meaning that something is evident to both the speaker and the hearer. However, in (3) clearly would have different connotation than obviously.
(3) A: »that this is a bottle of something
B:–UNCLEARWORD chocolate coated Brazil nuts
A: you think ?
C: clearly not chocolate coated Brazil nuts (Spoken BNC 2014)
Clearly has the meaning that something is self-evident to both the speaker and the hearer. Obviously in the same utterance would have a dismissive function since it is not self-evident that the hearer shares this opinion.

References

Alonso-Almeida, F. 2012. Sentential evidential adverbs and authorial stance in a corpus of English computing articles. Volumen Monográfico, 15-31.

Carretero, M. 2019. Evidentiality in adverbs of perceivability. The case of English manifestly, noticeably, patently and visibly. Functions of Language 26(3): 275-307.

Dendale, P. 2020. Are "modal adverbs" automatically modal markers? The case of French certainement with its epistemico-modal and its evidential use. Anuari de Filologia. Estudis de Lingüística 10: 39-76.

Love, R., Dembry, C., Hardie, A., Brezina,V. and T. McEnery. 2017. The Spoken BNC2014 - designing and building a spoken corpus of everyday conversations. International Journal of Corpus Linguistics 22 (3):319-344.

Rozumko, A. 2019. Evidential strategies in receiver-directed talk: The case of English inferential adverbs. Lingua 220: 1-16.

Simon-Vandenbergen, A.-M. and K. Aijmer. 2007. The semantic field of modal certainty. A corpus-based study of English adverbs. Berlin: Mouton de Gruyter.

Does ProtAnt identify prototypical corpus texts? A critique of Anthony & Baker (2015)

FULL PAPER

Stefan Th. Gries (UC Santa Barbara, US & JLU Giessen, Germany), Sebastian Hoffmann (University of Trier, Germany) & Nicholas I. Smith (University of Leicester, UK)

Anthony & Baker (2015, IJCL) introduces ProtAnt, a software tool that ranks the texts of a target corpus by their prototypicality based on the (relative) number of keywords that corpus contains, compared to a reference corpus. The rationale is that since keywords ”are distinctive of the target corpus as a whole along many possible different dimensions” (p. 278), a text with many such keywords will combine ”many of these different facets” (p. 279), establishing its role as ”a more central or typical text in that corpus” (p. 277). ProtAnt has been adopted in several studies aimed at downsizing a large corpus of candidate texts into a small pool of prototypical exemplars for close qualitative analysis (e.g. Bednarek & Caple, 2017; Kania, 2020; Candarli & Deignan, 2025).

Anthony & Baker seek to validate their approach with five experiments, one of which (experiment 5) involves checking essentially the opposite of identifying texts that are prototypical for a target corpus: the identification of outliers in a corpus. They compute for each register of the target corpus AmE06 (against BE06 as reference corpus) how a randomly-selected file from one of the other 14 registers gets ranked; as they themselves say, ”[i]deally ProtAnt should rate [such an] ’outlier’ file towards the bottom of the list of typical files” (p. 287).

In this paper, we present a series of results based on (i) replicating their experiment 5 but not just with 1 file for each register (i.e., their 15 tests), but with all files against each register (i.e., altogether 7,000 tests) and (ii) replicating while improving their experiment 5 by a) choosing a statistically better cut-off point for keywords to include and b) using a much more comprehensive operationalization of keyness (a 3-dimensional tuple of frequency, association, and dispersion, see Gries 2024: Section 5.4).

The results are extremely sobering. Even in the improved implementation of their experiment, the outlier file scores lowest in less than half of all cases: With their ranking statistic

of
normalized key tokens, the outlier file only scores lowest in 2,859/7,000 tests (40.8%); normalized key types, the outlier file only scores lowest in 3,258/7,000 tests (46.5%). Even more concerning, in 3% (212/7,000) of tests with key tokens and 3.4% (238/7,000) with key types, the outlier file is ranked as the most prototypical. More generally, 15.2% (1,067/7,000) (for tokens) and 14.6% (1,019/7,000) (for types) of outlier files from a random register get ranked in the upper half of files determined to be prototypical of the target register, and some of these results become even worse when their approach is actually improved as mentioned above.
We conclude by discussing why we think these results are obtained and how the ProtAnt algorithm is sub-optimal on both a theoretical level (its understanding of 'prototype') and a methodological level (its implementation).

References

Anthony, Laurence & Paul Baker. 2015. ProtAnt: A tool for analysing the prototypicality of texts. International Journal of Corpus Linguistics 20(3), 273-92.

Baker, Paul. 2009. The BE06 Corpus of British English and recent language change. International Journal of Corpus Linguistics 14(3), 312-37.

Bednarek, Monika & Helen Caple. 2017. The discourse of news values: How news organizations create newsworthiness. Oxford: Oxford University Press.

Candarli, D. and Deignan, A. 2025. Rhetorical moves in teachers' PowerPoint presentations: Variation across disciplines and school stages. Journal of English for Academic Purposes 76.

Gries, Stefan Th. 2024. Frequency, dispersion, association, and keyness: Revising and tupleizing corpus-linguistic measures. Amsterdam & Philadelphia: John Benjamins, pp. 321.

Kania, Ursula. 2020. Marriage for all ('Ehe fuer alle')?! A corpus-assisted discourse analysis of the marriage equality debate in Germany. Critical Discourse Studies 17(2), 138-55.

Potts, Amanda & Paul Baker. 2012. Does semantic tagging identify cultural change in British and American English? International Journal of Corpus Linguistics 17(3), 295-324.

Use of Corpora in Psycholinguistics: A Meta-Analysis of Academic Journals

WIP

Nina Dumrukcic (University of Cologne, Germany)

Despite notable overlaps between corpus linguistics and fields such as sociolinguistics or Critical Discourse Analysis, the integration of corpus methods within psycholinguistics remains limited. At major corpus linguistics conferences (e.g., ICAME, CL), psycholinguistic contributions are usually scarce, and few papers report experimental data. This meta-analysis investigates the extent to which corpora are used in psycholinguistic research by analyzing publications in peer-reviewed, open-access journals.
Corpora can serve a crucial role in experimental design—not only for measuring frequency,

but for identifying naturalistic sentence stimuli containing target lexical items. While corpora such as CHILDES are occasionally referenced, their systematic use in experimental design has not been thoroughly documented. Moreover, few studies provide replicable methodologies for integrating corpus data into psycholinguistic experiments, particularly in lexical access or sentence processing. As such, the potential of corpora to enhance ecological validity and stimulus control is often underutilized. Building on previous work on bridging the gap between corpus and psycholinguistics (Gilquin & Gries 2009), this study compiles a small corpus of academic papers published between 2015–2025 in four leading open-access psycholinguistic journals: Applied Psycholinguistics, Frontiers in Psychology (Psychology of Language), Glossa (Psycholinguistics), and PLOS One, yielding a dataset of 3,165 papers.

A webcrawler script (Strange 2025) using Python and Selenium was employed to automatically download papers from Frontiers in Psychology, while other articles were collected manually in batches. All documents were converted to plain text and merged for analysis. The resulting corpus was uploaded to SketchEngine and comprises approximately 11.8 million tokens, 8.5 million words, and over 421,000 sentences. Texts were tagged using part-of-speech and lemmatization tools, allowing for more nuanced searches related to lexical class and frequency distributions.

Preliminary findings suggest that while corpora are occasionally cited, their use for stimulus selection in experimental research remains relatively rare. References to corpora often appear in the context of frequency or using data from eye-tracking corpora such as GECO (Cop et al. 2017), with limited discussion of corpus-based sentence selection. This study examines the use of corpora for stimuli selection (Dumrukcic 2022) and evaluates frequency measures used by psycholinguists by utilizing lexical databases such as CELEX (Baayen 1995) or SUBTLEX (Brysbaert 2009) to determine lexical salience as opposed to using corpus linguistic frequency measures.

This meta-analysis offers insight into methodological trends and encourages greater interdisciplinary integration. Further discussion addresses the benefits of corpus-driven stimulus design and proposes future directions for improving the synergy between experimental and corpus approaches in psycholinguistic research.

References

Baayen, R. H., Piepenbrock, R., & van Rijn, H. (1995). The CELEX lexical database (CD-ROM). Linguistic Data Consortium, University of Pennsylvania.

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. Behavior Research Methods, 41(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Cop, U., Dirix, N., Drieghe, D., & Duyck, W. (2017). Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. Behavior Research Methods, 49, 602–615. https://doi.org/10.3758/s13428-016-0734-0

Dumrukcic, N. (2022). Translanguaging and the bilingual brain: A mixed methods approach to word-formation and language processing. Walter de Gruyter GmbH & Co KG.

Gilquin, G., & Gries, S. T. (2009). Corpora and experimental methods: A state-of-the-art review. Corpus Linguistics and Linguistic Theory, 5(1), 1–26. https://doi.org/10.1515/CLLT.2009.001

Strange, C. (2025). webcrawler frontiers in [GitLab repository]. GitLab. https://gitlab.com/codybstrange/webcrawler-frontiers-in

"A levels and stuff like that"/"Aunque yo no esté por ella, ella sigue conmigo, bla bla bla tía"/. General Extenders in Spanish and English Teen Talk: A Corpus-Based Contrastive Study. FULL PAPER

Ignacio M. Palacios Martinez (University of Santiago de Compostela, Spain)

Teenagers are often seen as agents of linguistic change (Eckert 1988, Kerswill 1996, Stenström et al. 2002, Tagliamonte 2016), regularly introducing new forms and uses that may eventually become part of the general structure of the language. This paper focuses on the analysis of a group of English and Spanish expressions described in the literature as general extenders (GEs), vague category identifiers, or utterance-final coordinator tags in English (Overstreet 1999; Cheshire 2007; Tagliamonte & Denis 2010; Overstreet & Yule 2021; Brinton 2024), and as elementos de final de serie enumerativa and fórmulas generalizadoras in Spanish (Cortés 2006; Rodríguez 2015; Borreguero 2022). These include forms such as and stuff, or whatever in English, and y todo, y eso, y tal, o algo (así) in Spanish. The interest in GEs lies not only in their syntax and semantics, but also in their pragmatics. Contrary to the expectation that GEs are merely markers of vagueness (Chanell 1994, Cutting 2007, Ruzaitè 2007), they often perform additional intersubjective functions.

The analysis here draws on data taken from COLT, DCPSE, and BNC2014 for English, and COLAm, and CORPES XXI for Spanish.

Four main hypotheses are tested:

1. GEs will be more common in the language of teenagers than in that of adults, since they are, in principle, more frequent in informal speech, which is favoured by teen speakers.
2. GEs in teen talk will convey pragmatic meanings that may differ from those typically found in adult usage.
3. GEs, as some scholars claim (Overstreet 1999; Cheshire 2007), show evidence of processes of change typically associated with grammaticalization (Hopper & Traugott 1993; Heine & Kuteva 2002). At the same time, Tagliamonte & Denis (2010) and Tagliamonte (2016) argue that these developments may be more accurately described as instances of lexical replacement, while Brinton (2024) maintains that they are better accounted for within a constructionist framework, at least with regard to the analysis of what-GEs in the history of English.
4. These tendencies will be observable in both Spanish and English, with no major cross-

linguistic differences expected.
Initial results show that and stuff and or something have increased exponentially among both young and adult speakers in recent years, becoming the most common general extenders in the language of British teenagers. However, and things, or something and and that are more frequent in adult speech. In Spanish, y todo and y tal (cual) are the most recurrent among teenagers, while y tal (cual) and etcétera, etcétera are dominant in adult usage. Interestingly, while the use of GEs in Spanish is clearly more frequent among younger speakers, in English the opposite tendency is apparently identified. The interpersonal function of solidarity or complicity is particularly prominent in the language of teenagers. Finally, features typically associated with grammaticalisation and pragmaticalisation are observed in the teen talk of both languages.
The conclusions of this study aim to contribute to a deeper understanding of GEs and of adolescent language use in both Spanish and English.

References

Borreguero Zuloaga, M. (2022). General extenders in Spanish interactions: Frequent forms, pragmatic functions y todo eso. Anuari de Filologia. Estudis de Lingüística 12, 155-187. https://doi.org/10.1344/AFEL2022.12.8.

Brinton, L. (2024). The rise of What-general extenders in English. Journal of Historical Pragmatics, 25(1), 104-136.

Cheshire, J. (2007). Discourse variation, grammaticalisation and stuff like that. Journal of Sociolinguistics, 11(2), 155-193.

Cortés Rodríguez, L. (2006). Los elementos de final de serie enumerativa del tipo y todo eso, o cosas así, y tal, etc. Perspectiva interactiva. Boletín de Lingüística XVIII/26, 102-129.

Gille, J. & Häggkvist. C. (2006). Los niveles del diálogo y los apéndices conversacionales. In J. Falk, J. Gille, & F. Wachtmeister Bermúdez (Ed.), Discurso, interacción e identidad (pp. 65-80). Stockholm University.

Heine, B. & Kuteva, T. (2002). World lexicon of grammaticalization. Cambridge: University Press.

Hopper, P. J. & Closs Traugott, E. (1993). Grammaticalization. Cambridge University Press.

Overstreet, M. (1999). Whales, candlelight, and stuff like that: General extenders in English discourse. Oxford University Press.

Overstreet, M. & Yule, G. (2021). General extenders. The forms and functions of a new linguistic category. Cambridge University Press.

Rodríguez, J. (2015). General extender use in spoken Peninsular Spanish: Metapragmatic awareness and pedagogical implications. Journal of Spanish Language Teaching, 2(1), 1-17.

Tagliamonte, S. A.& Denis, D. (2010). The stuff of change: General extenders in Toronto, Canada. Journal of English Linguistics, 38(4), 335-368. https://doi.org/10.1177/0075424210367484.

Tagliamonte, S. A. (2016). Teen talk: The language of adolescents. Cambridge University Press.

Corpus-Based Evidence of Creole Influence in Tobagonian English (ICE T&T)

FULL PAPER

Noumidia Allouch (Christian Albrecht University of Kiel, Germany)

Despite extensive ICE data, the Caribbean remains underrepresented, with ICE-Jamaica long the sole completed corpus until the recent addition of ICE-T&T for Trinidad and Tobago (Deuber, 2014, 13). Although ICE-T&T has been studied (notably Deuber, 2014), research has focused on Trinidadian English due to its larger speaker base and data representation. Though politically unified – sharing English as the official language – the islands' diverging colonial histories have resulted in different sociolinguistic profiles (Westphal et al., 2023, 4): In Trinidad, Standard English and a mesolectal English-based Creole coexist in diglossic distribution (Winford, 1985). Tobago's situation includes a basilectal creole, preserved by its long-standing village-based society (Youssef & James, 2004, 513). According to Youssef (2010, 67; 2004, 44), the creole marks identity, solidarity, emotion and humour whereas the standard is used in contexts of education, religion and officialdom. Linguistic research on Tobago has centred on the creole, not the standard (James & Youssef, 2004; Youssef & James, 2004). Drawing on the Tobagonian subcorpus of ICE-T&T, this study examines Standard Tobagonian English to address this research gap. It is hypothesized that the Tobagonian subcorpus exhibits a higher density of mesolectal structures than its Trinidadian counterpart, which is why creolized features known as creolisms are of particular interest (Allsopp, 1996; Mair, 2002). By analysing variation in progressive forms – standard progressive and creolized zero-copula progressive (e.g. They Ø coming from England), a salient feature of the mesolect (Youssef, 2011, 201–202) – the paper advances understanding of creole influence on Tobagonian English grammar.

Focusing on Tobagonian speakers yields a significantly smaller data set of 32 texts. Although limited in register and speaker variation – private conversations among educated Tobagonians – the subcorpus enables a more fine-grained analysis. Aside from aspectual function, this paper examines the role of speech topics in shifts between creole and standard, which has received less attention than register and social factors (e.g. Youssef, 2011). While conversation topics must be considered within a multifaceted communicative context (Deuber, 2014, 114), their influence appears significant as evidence suggests (Youssef, 1993). The study addresses the following questions:

1) How is creole influence reflected in the use of progressive forms?

2) How does progressive function (e.g. habituality) influence choice and distribution of specific progressive forms?

3) Considering Youssef's (2004, 44) account on creole and standard use, to what extent does the type of topic (e.g. identity or education) affect progressive form variation?

Using MaxQDA for topic coding alongside traditional corpus-linguistic tools, the progressive forms are examined with regards to frequency and contextual use to determine aspectual function. Topics are categorized bottom-up into suitable types to examine their interaction with grammatical form and function. The results are expected to reveal an extended use of

the progressive, paralleling findings from other Outer Circle varieties like Nigerian English or Black South African English (e.g. Gut & Fuchs, 2013; van Rooy, 2006; 2014). Based on Youssef's account, it is investigated whether the zero-copula progressive form will correlate with topics associated with the creole.

References

Allsopp, R. (1996). Dictionary of Caribbean English usage (1. publ). Oxford Univ. Press.

Deuber, D. (2014). English in the Caribbean: Variation, style and standards in Jamaica and Trinidad. Studies in English Language. Cambridge University Press.

Gut, U., & Fuchs, R. (2013). Progressive Aspect in Nigerian English. Journal of English Linguistics, 41(3), 243–267. https://doi.org/10.1177/0075424213492799

James, W., & Youssef, V. (2004). The creoles of Trinidad and Tobago: morphology and syntax. In B. Kortmann & E. W. Schneider (Eds.), A Handbook of Varieties of English (pp. 454–481). De Gruyter.

Mair, C. (2002). Creolisms in an emerging standard. English World-Wide. A Journal of Varieties of English, 23(1), 31–58. https://doi.org/10.1075/eww.23.1.03mai

van Rooy, B. (2006). The extension of the progressive aspect in Black South African English. World Englishes, 25(1), 37–64. https://doi.org/10.1111/j.0083-2919.2006.00446.x

van Rooy, B. (2014). Progressive aspect and stative verbs in Outer Circle varieties. World Englishes, 33(2), 157–172. https://doi.org/10.1111/weng.12079

Westphal, M., Deuber, D., & Ka Man Lau, D. (2023). Manual for the Trinidad and Tobago component (ICE-T&T), 1–17. https://www.uni-muenster.de/ICE/ICE-T_T.html

Winford, D. (1985). The Concept of "Diglossia" in Caribbean Creole Situations. Language in Society, 14(3), 345–356. https://www.jstor.org/stable/4167664?seq=1

Youssef, V. (1993). Children's Linguistic Choices: Audience Design and Societal Norms. Language in Society, 22(2), 257–274. http://www.jstor.org/stable/4168433

Youssef, V. (2004). 'Is English we speaking': Trinbagonian in the twenty-first century. English Today, 20(4), 42–49. https://doi.org/10.1017/S0266078404004080

Youssef, V. (2010). Varilingualism: A term for 21st century language acquisition contexts. Education Et Sociétés Plurilingues(28), 65–76.

Youssef, V. (2011). The varilingual repertoire of Tobagonian speakers. In Variation in the Caribbean (pp. 191–206). John Benjamins. https://www.jbe-platform.com/content/books/9789027287397-cll.37.12you

Youssef, V., & James, W. (2004). The creoles of Trinidad and Tobago: phonology. In B. Kortmann & E. W. Schneider (Eds.), A Handbook of Varieties of English (pp. 508–524). De Gruyter.

”Please do me the favor of not denying it!” - How conversational patterns in conflict talk shape fictional characters on television

FULL PAPER

Christian Hoffmann (University of Augsburg, Germany)

Although natural language data on conflict talk is extremely hard to gather, during the last four decades scholars have shed much light on the structural properties and conversational organisation of such “backstage language behaviour” (Goffman 1959). The resulting heap of studies have mainly used conversational data from a series of private or institutional settings to study conflicts, ranging from arguments exchanged between family members at the dinner table (Vuchinich 1990; Boxer 2008; Clift & Pino 2020) to conflict resolution strategies employed in adversative interactions between employees in companies (Rønneberg & Svennevig, 2010; Boehringer & Karl 2015; Hall & Butler 2017) or critical debates between politicians, journalists or game show participants (Thornborrow 2000; Lorenzo-Dus 2009, Hutchby 2013; Sinkeviciute 2015). Similarly, research on impoliteness has frequently drawn on selective examples from exploitative television shows or fictional Hollywood movies to discuss the use and conceptual limits of impoliteness frameworks (cf. Culpeper 2005; Culpeper & Holmes 2013). Yet, systematic corpus-based analyses tracking (different features of) conflict talk in (and across different) fictional TV shows are extremely rare.
To close this research gap, this paper adapts frameworks by Vuchinich (1990) and Bousfield (2007) to detect and classify the structural and functional properties of 65 different conflict talk sequences from twenty contemporary fictional comedy and drama TV shows produced in the USA. Manual coding of the corpus includes the categorisation of participant types (gender, age, character role and type), types of offending events, response sequences, defensive counter strategies and conflict terminations. The findings not only provide quantitative insights into the narrative design of conflict talk across TV shows in both genres but also reveal how conflict patterns tend to correlate with specific character attributes, e.g. age, gender, status, or narrative role. In addition, the talk showcases the main results of a small qualitative analysis of 20 conflict talk sequences involving the two main characters of Walter and Skyler White from the widely acclaimed TV series Breaking Bad. The incremental change of Walter’s conversational demeanour across the different seasons of the show subtly reflects his transformation “from a mild-mannered teacher to a calculating killer” (Wakeman 2018, 217). This illustrates how gradual modifications to the conversational design of conflict talk across different episodes constitutes an essential strategy for characterisation in (post-)modern fictional television (Culpeper 2001).

References

Boehringer, D., & Karl, U. (2015). “Do You want to negotiate with Me?”– Avoiding and dealing with

conflicts arising in conversations with the young unemployed. Social Work & Society, 13(1), pp. 1-17.

Bousfield, D. (2007). Beginnings, middles and ends: A biopsy of the dynamics of impolite exchanges. Journal of Pragmatics, 39(12), pp. 2185-2216.

Boxer, D. (2008). Face-to-face in the family domain. In: Diane Boxer (ed.) Applying Sociolinguistics (pp. 21-46). Amsterdam: Benjamins.

Clift, R., & Pino, M. (2020). Turning the tables: Objecting to conduct in conflict talk. Research on Language and Social Interaction, 53(4), pp. 463-480.

Culpeper, J., & Holmes, O. (2013). (Im) politeness and exploitative TV in Britain and North America: the X factor and American idol. In: Nuria Lorenzo-Dus & Pilar Garcés-Conejos Blitvich, Real talk: Reality television and discourse analysis in action (pp. 169-198). Basingstoke: Palgrave Macmillan.

Culpeper, J. (2001). Language and Characterisation. People in Plays and Other Texts. London: Routledge.

Goffman, E. (1956). The Presentation of Self in Everyday Life. New York: Doubleday.

Hall, J. K., & Butler, E. R. (2017). The shifting role of a document in managing conflict and shaping the outcome of a small group meeting. Text & Talk, 37(5), pp. 615-638.

Hutchby, I. (2013). Confrontation talk: Arguments, asymmetries, and power on talk radio. London: Routledge.

Lorenzo-Dus, N. (2009). "You're barking mad, I'm out": Impoliteness and broadcast talk. Journal of Politeness Research: Language, Behavior, Culture, 5(2), pp. 159-187

Rønneberg, K., & Svennevig, J. (2010). Declining to help: Rejections in service requests to the police. Discourse & Communication,4(3), pp. 279-305.

Sinkeviciute, V. (2015). "There's definitely gonna be some serious carnage in this house" or how to be genuinely impolite in Big Brother UK. Journal of Language Aggression and Conflict, 3(2), pp. 317-348.

Thornborrow, J. (2000). The construction of conflicting accounts in public participation TV. Language in Society, 29(3), pp. 357-377.

Vuchinich, S. (1990). The sequential organization of closing in verbal family conflict. In: Allen D. Grimshaw (ed.), Conflict Talk: Sociolinguistic Investigations of Arguments and Conversations (pp. 118-138). Cambridge: Cambridge University Press.

Wakeman, S. (2018). The "'one who knocks" and the "one who waits": Gendered violence in Breaking Bad. Crime, Media, Culture, 14.2, pp. 213-228.

FROM BEAT TO SPEECH: TEACHING ENGLISH RHYTHM THROUGH MUSICAL TRAINING

FULL PAPER

Giedrė Balčytytė (Vilnius University, Lithuania)

In contemporary communicative settings, fluency has become a key indicator of successful language use, extending far beyond accurate articulation. Prosody, and in particular rhythm, underpins this fluency by shaping both temporal and intonational patterns of

connected speech. The English rhythm, commonly characterised as stress-timed, has long been debated (Roach, 2009), yet for pedagogical and perceptual purposes this distinction remains useful, especially for learners whose first languages are syllable-timed or mora-timed. Rhythm directly influences connected-speech phenomena such as reduction, linking and elision, which together create the impression of natural fluency (Crystal, 2012). Recent research confirms that rhythm perception is shaped by both native-language rhythmic typology and individual beat-perception ability (Smit et al., 2024), reinforcing the need to address rhythm explicitly in second-language phonetic instruction.

This study examines whether short-term, music-based phonetic training can enhance the rhythmical and prosodic fluency of adolescent learners of English as a Foreign Language (EFL). Building on neurocognitive research demonstrating shared temporal and processing mechanisms in music and speech (Patel, 2008; Tierney & Kraus, 2013), the research hypothesised that rhythm-focused musical practice would improve learners' control of prosodic timing and connected-speech patterns.

47 EFL learners aged 14-17 representing 18 nationalities and three rhythm-type language backgrounds (stress-timed, syllable-timed and mora-timed) participated in the experiment. Participants received ten minutes of rhythm-based training daily over two weeks during regular EFL instruction. The training included rhythmic chanting, syllable clapping and sung intonation patterns, designed to reinforce stress placement and timing. Musical aptitude was assessed through a short perception test adapted from Gordon's Advanced Measures of Music Audiation (1989).

The recordings formed a small spoken learner corpus designed for comparative analysis of rhythm and prosody before and after the intervention. Speech samples were recorded before and after the training using a standard reading passage and a brief spontaneous narrative task. The corpus was analysed acoustically in Praat with attention to speech rate, inter-stress timing and the duration of vocalic and consonantal intervals. Measures of connected-speech fluency, such as the use of linking, weak forms and reduction were also examined through auditory analysis. Descriptive and inferential statistics were applied to compare pre- and post-training performance, using R for basic significance testing. The post-intervention results indicated measurable improvement in rhythm regularity, smoother timing between stressed syllables and more natural connected-speech patterns. Learners with higher musical aptitude showed stronger rhythmic consistency and enhanced temporal fluency, supporting the view that sensitivity to musical rhythm facilitates prosodic development in a second language.

The findings confirm that rhythm training grounded in musical patterns enhances both the perceptual and productive aspects of English prosody. Beyond phonological benefits, the approach fosters cognitive engagement and supports learners' sense of temporal fluency. The study contributes to growing evidence that integrating music into phonetic pedagogy provides an effective and humanly engaging route to developing prosodic competence in English.

References

Crystal, D. (2012). The Cambridge encyclopedia of the English language (2nd ed.). Cambridge University Press.

Gordon, E. E. (1989). Advanced measures of music audiation. GIA Publications.

Patel, A. D. (2008). Music, language and the brain. Oxford University Press.

Roach, P. (2009). English phonetics and phonology (4th ed.). Cambridge University Press.

Smit, E. A., & Rathcke, T. (2024). The role of native language and beat perception ability in the perception of speech rhythm. Psychonomic Bulletin & Review.

Tierney, A., & Kraus, N. (2013). The ability to move to a beat is linked to the consistency of neural responses to sound. Journal of Neuroscience, 33(38), 14981–14988.

RHINE

MOSELLE

Author Index

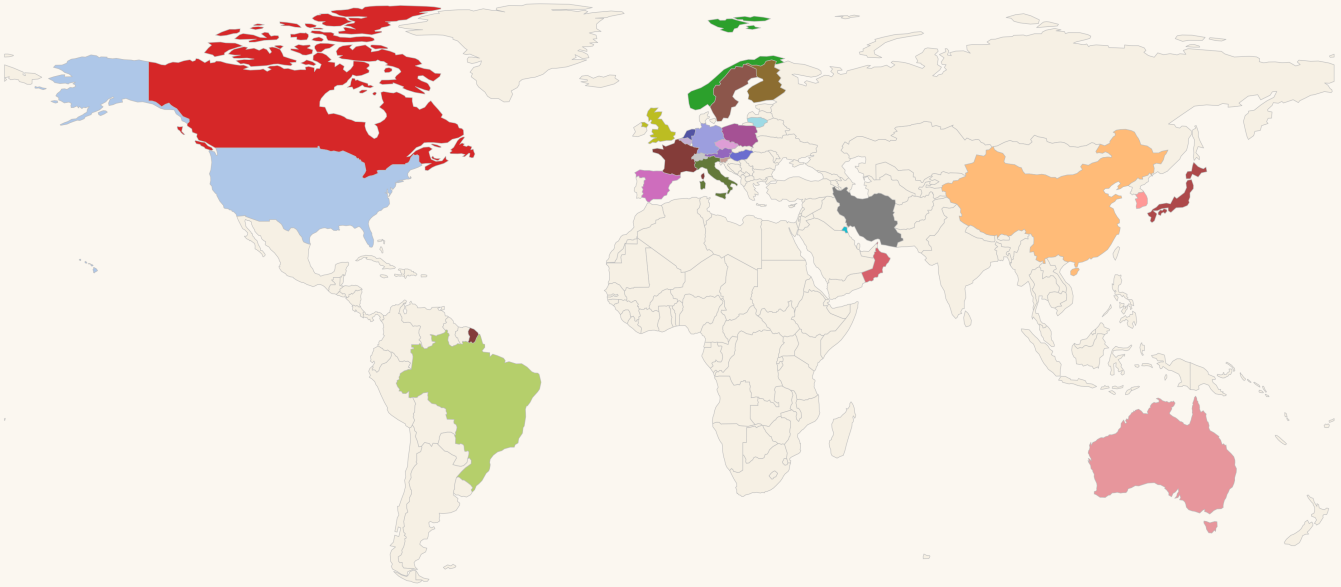
Page numbers refer to the abstract entry for each author. Authors appearing in workshop presentations are indexed under their respective workshop section.

Author Index

COUNTRY AND INSTITUTION REPRESENTATION

Global – Europe – Germany

WORLD



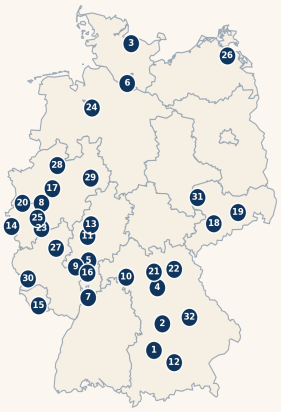
Countries Represented

Australia	Canada	France	Italy	Maldives	Poland	Sweden
Austria	China	Germany	Japan	Netherlands	Slovenia	Switzerland
Belgium	Czech Republic	Hungary	Kuwait	Norway	South Korea	UK
Brazil	Finland	Iran	Lithuania	Oman	Spain	USA

EUROPE

Austria	Hungary	Slovenia
Belgium	Italy	Spain
Czech Republic	Lithuania	Sweden
Finland	Netherlands	Switzerland
France	Norway	United Kingdom
Germany	Poland	

GERMANY



1. Augsburg University
2. Catholic University of Eichstätt-Ingolstadt
3. Christian Albrecht University of Kiel
4. FAU Erlangen-Nürnberg
5. Goethe University, Frankfurt am Main
6. Hamburg University
7. Heidelberg University
8. Heinrich Heine University Düsseldorf
9. JGU Mainz
10. Julius Maximilians University Würzburg
11. Justus Liebig University Giessen
12. LMU Munich
13. Philipps-University, Marburg
14. RWTH Aachen University
15. Saarland University
16. TU Darmstadt
17. TU Dortmund
18. Technische Universität Chemnitz
19. Technische Universität Dresden
20. University of Applied Sciences Düsseldorf
21. University of Bamberg
22. University of Bayreuth
23. University of Bonn
24. University of Bremen
25. University of Cologne
26. University of Greifswald
27. University of Koblenz
28. University of Münster
29. University of Paderborn
30. University of Trier
31. Universität Leipzig
32. Unversity of Regensburg

ICAME 47

International Computer Archive
of Modern and Medieval English

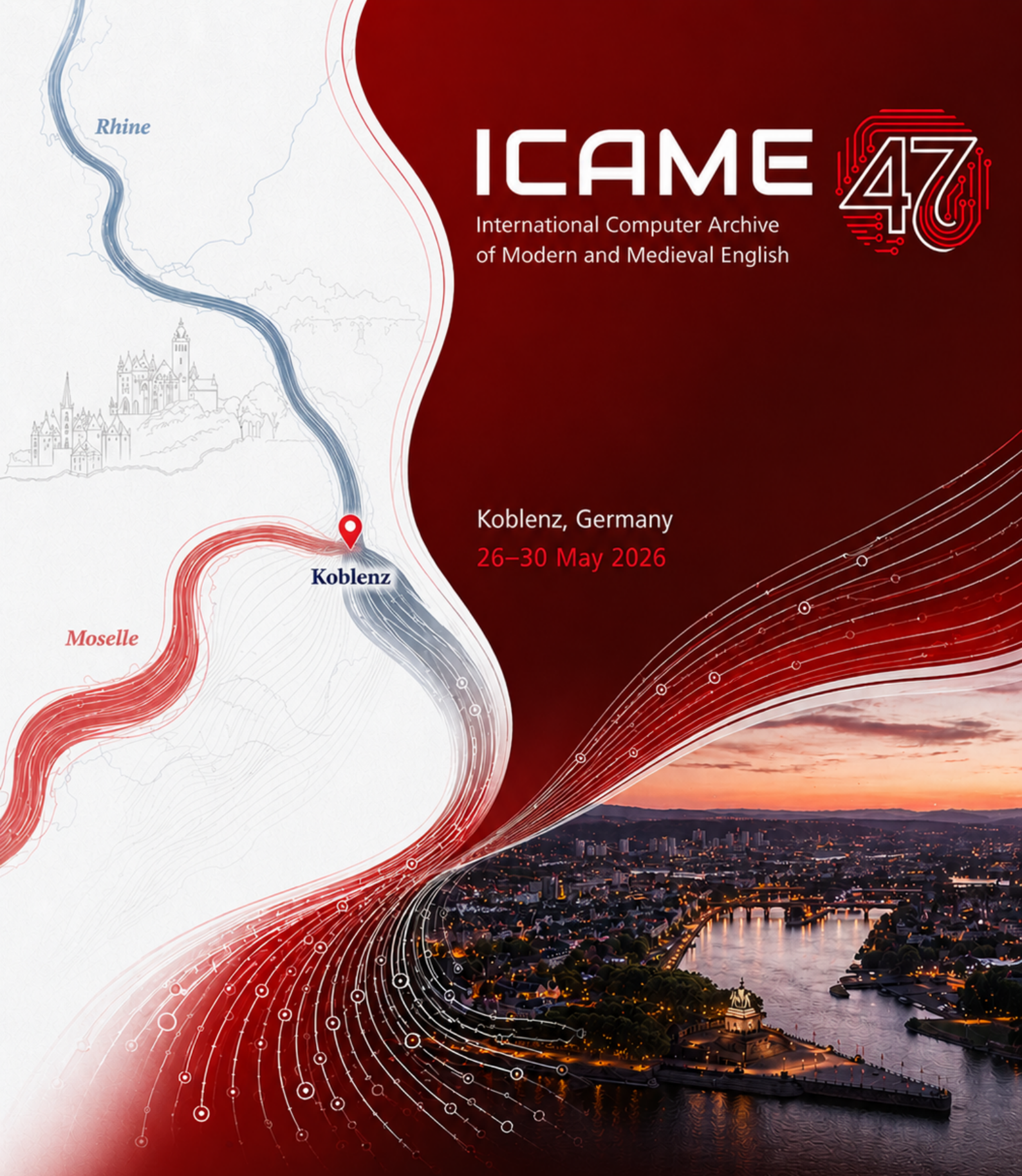
Rhine

Koblenz

Moselle

Koblenz, Germany

26–30 May 2026



A CONFLUENCE OF CORPUS RESEARCH IN THE AGE OF AI

HOSTED BY

university of koblenz
Arts / Humanities

SPONSORED BY

DFG Deutsche Forschungsgemeinschaft
German Research Foundation

insiders