AACL 2026

UNIVERSITY OF FLORIDA

Welcome!

On behalf of the organizing committee, it is my pleasure to welcome you to the 2026 conference of the *American Association for Corpus Linguistics*. We hope you will enjoy the many intriguing oral and poster presentations, and we are excited to share our beautiful campus with you.

I would like to thank the organizing committee for undertaking this enormous task. Jessi Aaron, Firoz Ahmed, Alaa Albeladi, Helene Blondeau, Kathryn Conger, Jamie Garner, Jessica Heady, Zoey Liu, Maria Laura Mecias, Akindele Olalekan, Valeria Pagliai, Jaylen Parker, Emiliana Pulido, Katrina Smith, Haiyin Yang, and Minghao Zheng, you all rock!

We also would like to express our gratitude to our sponsors: the Alachua County Meeting Advantage Program; the William and Grace Dial Center for Speech and Communication Studies at UF; the UF Center for European Studies; the UF Center for Humanities and the Public Sphere; the UF Center for Latin American Studies; the UF College of Liberal Arts and Sciences; the UF Department of English; the UF Department for Spanish and Portuguese Studies; and the UF Linguistics Department. Special thanks go to Lynne Palmer (Assistant Director, UF Reitz Union Event Services) and to Kelli Granade (Key Administrator, UF Linguistics Department) for their tireless support. And of course, we are grateful to everyone who reviewed for this conference.

This program serves as a complete record for the conference. You can find the schedule for oral presentations, a list of all poster presentations, our plenary speakers, and all presentations and poster abstracts here. For quick reference, scan this QR code, which takes you to the online version of the schedule; there, all presentations and poster titles are hyperlinked to their abstracts. The same QR code is also on the back of your name tags for easy access.

If there is anything you need while you are here, do not hesitate to talk to me or anyone wearing an AACL2026 T-shirt. We hope you will have a great time in the Gator Nation!

Warmly,
Steffi

Stefanie Wulff
AACL2026 Organizing Chair

Table of Contents

Schedule

Saturday, 4/18/2026

Time	@Arredondo Café	@Chamber	@G310	@G320	@G325
8am-10:30am	Registration				
8:30-9am		Opening Remarks			
9-10:30am		Plenary Talk 1: COWS-L2H: a corpus of L2 and heritage Spanish at the intersection of computational linguistics and language instruction (*Kenji Sagae*) (introduction: *Zoey Liu*)			
10:30-11am	Coffee Break				
		Session chair: *Ratree Wayland*	Session chair: *Tove Larsson*	Session chair: *Ryan Ka Yau Lai*	Session chair: *Shelley Staples*
11-11:25am		# 107 Linguistic variation in identity-first vs. person-first language among autism community stakeholders in a reddit corpus (*Alyssa Lowry*, Earl Brown)	# 150 Linking adverbials in academic writing: insights from L1 and L2 student corpora (*Duong Nguyen*)	# 110 Self-mention in discussion sections: a corpus-based comparison of single and co-authored research papers from applied linguistics journals (*Juan Rostrán Valle*)	# 18 From shielding lives to growing futures: a corpus-based study of multimodal metaphors in Korean insurance posters (*Ebru Turker*)
11:30-11:55am		# 136 Neutralization and cue-weighting of sibilant contrasts in Parkinson's speech: a cross-corpus study (*Firoz Ahmed*, Ratree Wayland)	# 71 Prompt effects on L2 English writing (Henrik Kaatari, *Taehyeong Kim*, Tove Larsson, Ying Wang, Pia Sundqvist)		
12-2pm	Poster session (@Rion Ballroom!)				
		Session chair: *Irina Burukina*	Session chair: *Lizzy Hanks*	Session chair: *Juan Rostran Valle*	Session chair: *Ebru Turker*
2-2:25pm		# 50 Register as a predictor of argument structure construction use (Jesse Egbert, *Hakyung Sung*)	# 69 Investigating EFL student writing in the CAWESA: an additive multidimensional analysis (*Wesley Acorinti*,	# 158 Uncovering cross-linguistic register variation with a shared MD model of English and Chinese *(Shangyu Jiang*)	# 164 Corpus approaches to media, language, and the politics of recognition (*Caroline Scheuer Neves*)

			Marine Matte, Larissa Goulart)		
2:30-2:55pm		# 48 Corpus-based identification of Mandarin dative construction prototypes using random forests (Shengyu Liao, *Stefan Th. Gries*, Stefanie Wulff)	# 66 Additive discourse markers in academic writing: a corpus-based comparative study (*Abdulwahab Alshehri*)	# 141 Bootstrapping tupleised statistics to compare construction inventories: the case of Cantonese and Mandarin particle frames (*Ryan Ka Yau Lai*)	# 104 From reputation to survival: student discourses of university life in reviews and reddit (*Savannah T. Brown*, Jack A. Hardy)
3-3:25pm		# 116 Here's when non-alternating examples can be included in alternation research: with the right predictive modeling approach (*Stefan Th. Gries*, Nina S. Funke)		# 134 A corpus approach to constructional variation and category change in Portuguese (*Chad Howe*)	# 135 Linguistic variation and team performance: a corpus study of Brazilian football fans (*Alexcia Boothe*)
3:30-4pm	Coffee Break				
		Session chair: *Caroline Scheuer-Neves*	Session chair: *Wesley Acorinti*	Session chair: *Emi Pulido*	Session chair: *Kathryn Conger*
4-4:25pm		# 52 From community-oriented documentation to a socio-linguistically diverse corpus (*Irina Burukina*, Polina Pleshak, Maria Polinsky)	# 41 Do first-year composition assignments align with the linguistic and situational demands of disciplinary writing? (Larissa Goulart, *Elizabeth Hanks*)	#122 Beyond the news: genre-diverse annotations for bridging anaphora in English (*Lauren Levine*, Amir Zeldes)	# 90 Who stops bleeding?: a diachronic study of amenorrheic discourses in The Lancet (Veronica Ma, *Jack A. Hardy*, Paige Crowl)
4:30-4:55pm		# 151 Surveying native speakers' real time register use to design a representative corpus (*Andrea Flinn*)	# 58 Mapping disciplinary writing development with multi-dimensional analysis (*Haiyin Yang*, Stefanie Wulff)	# 77 Language of distress: machine learning and corpus-linguistic analysis of depression, PTSD, and anxiety in online communities (*Youngmeen Kim*, Ute Römer-Barron)	# 21 Interdisciplinary applied corpus linguistics: a case study of a promising partnership and some questions for the field (*Shannon Fitzsimmons-Doolan*, Jennifer Beseres Pollack)
5:15-6:45pm		Plenary Talk 2: Community corpora for linguistics science: studies from a project about history, culture and people (*Sali Tagliamonte*) (introduction: *Helene Blondeau*)			

7-9pm	Reception (@Rion Ballroom!)				

Sunday, 4/19/2026

Time	@Arredondo Café	@Chamber	@G310	@G320	@G325
8am-10:30am	Registration				
9-10:30am		Plenary Talk 3: A picture is worth 1000 words? A multimodal investigation of an image-tagged corpus (*Paul Baker*) (introduction: *Steffi Wulff*)			
10:30-11am	Coffee Break				
		Session chair: *Febriana Lestari*	Session chair: *Marianna Grachova*	Session chair: *Ute Roemer-Barron*	Session chair: *Maha Al-Harthi*
11-11:25am		#12 Xenophobic representations targeting China on "X" during the COVID-19 pandemic (*Cicero Soares da Silva*)	# 33 Exploring the humanlikeness of AI-generated texts through register alignment: a corpus-based study of ChatGPT (*Yağmur Demir*)	# 105 Enhancing first-year writing instruction at Hispanic serving institutions with corpus-informed pedagogy: resources, strategies and responsiveness (*Anh Dang*, Shelley Staples, Randi Reppen)	# 54 "I can't tell bad stories in Spanish:" comparing the effect of topic valence on lexis in oral vs. written narrations from L2 Spanish learners (*Andrea Hernandez*, Sophia Minillo, Claudia Helena Sanchez-Gutierrez, Paloma Fernandez-Mira)
11:30-11:55am		# 17 Variability within Stability: A Novel Key Keyword Multidimensional Analysis of the Diachronic Coverage Change in China Daily's International Media Reporting on China (2017-2024) (*Chenghui Wu*)		# 132 The relationship between L2 Spanish written narrative retellings and features of lexicogrammatical use (*Hana Dussan*, Kristopher Kyle, Mery Díez-Ortega, Carla Consolini)	# 85 L2 writing in Ghanaian high-stakes exams: a register-functional analysis (*Bernard Cassie*, Kwaku Osei-Tutu, Shelley Staples)
12-1:30pm	Lunch				
		Session chair: *Daniel Keller*	Session chair: *Valeria Pagliai*	Session chair: *Yigit Savuran*	Session chair: *Tsukuru Kamiyama*
1:30-1:55pm		# 26 From rates of occurrence to prototypicality:	# 84 Generative AI for academic writing: comparing few-shot	# 22 Teaching algebra online: patterns of teacher talk and their	#112 A lexicogrammatical analysis of a phrasal

		mapping the distinctive features of conversation (*Elizabeth Hanks*)	and zero-shot in table-to-text generation (*Yiwen Zheng*, Daniel Dixon)	relationship to student outcomes (*Zhihui Fang*, Zifeng Liu, Guo Rui, Wanli Xing, Ning Mao)	complexity feature (*Elizabeth Meyr*, Brett Hashimoto)
2-2:25pm		# 83 What's salient? Genre matters! (*Amir Zeldes*)	# 68 Leveraging GenAI for proficiency-aligned feedback on L2 learner writing (*Shuyuan Tu*)	# 55 From learner corpus to classroom: how can we translate usage-based SLA findings into pedagogical practice? (*Ute Römer-Barron*)	# 111 Measuring lexical complexity in L2-Korean writing through a morpheme-aware approach (*Hakyung Sung*, Gyu-Ho Shin Shin)
2:30-2:55pm		# 100 Text-internal register shifts in web texts - a cross-linguistic approach (*Veronika Laippala*, Alireza Razzaghi, Erik Henriksson)	# 73 GPT-based assisted editing of pre-publication academic writing: an additive multi-dimensional analysis (*Rogerio Yamada*)		# 139 Lexical complexity as a lens on varied academic experiences in L1 and L2 postgraduate writing (*Maha Al-Harthi*)
3-3:30pm	Coffee Break				
		Session chair: *Kathryn Conger*	Session chair: *Rogerio Yamada*	Session chair: *Hana Dussan*	Session chair: *Elizabeth Meyr*
3:30-3:55pm		# 125 Communicative co-occurrence in conversation: intra-text patterns of communicative purpose and topic across conversational discourse units in BNC Spoken (*Daniel Keller*, Marianna Gracheva)	# 40 LLMs and ideological priming: a case study of immigration discourse (*Tony Berber Sardinha*, Anderson Avila, Maria Claudia Nunes Delfino, Marilisa Shimazumi, Deise Prina Dutra, Paula Tavares Pinto, Ana Bocorny, Carlos Kauffmann, Patricia Bértoli, Mirella Whiteman, Marcos Roberto de Oliveira, Rogerio Yamada, Leandro Tessler)	# 44 Profiling the L2 Turkish lexicon: a learner corpus approach to CEFR-based vocabulary lists (*Yigit Savuran*, Stefanie Wulff)	#115 Expanding the construct of grammatical complexity: the case for grammatical diversity in writing (Christian Holmberg Sjöling, *Taehyeong Kim*)
4-4:25pm		# 127 Audiences and discourse types in fiction: a register analysis of children's and adult literature (*Marianna Gracheva*, Michaela Mahlberg)	# 76 Lexical and syntactic predictors of human-judged readability: an interpretable machine learning analysis of main effects and genre interactions (*Youngmeen Kim*)	# 49 Stability and usability of word lists based on forms, lemmas, flemmas, and word families in a specialized corpus (*Brett Hashimoto*, Elizabeth Hanks, Kyra Larsen, Jesse Egbert)	# 113 Developmental trajectories of noun phrase complexity in L2 English academic writing: frequency and lexical realizations (*Taehyeong Kim*)
4:30-4:55pm		# 155 Disciplinary register variation across the 'ages': From undergraduate	# 119 Lexical and lexico-grammatical predictors of heritage Spanish bilinguals'	# 97 A corpus-based approach to a CEFR-aligned Korean vocabulary profile	# 32 Text-level grammatical complexity revisited: Are there additional

		writer to researcher-in-training to disciplinary expert (*Bethany Gray*, Febriana Lestari, Duong Nguyen, Kimberly Becker)	productive proficiency: a replication of Kyle and Eguchi (2023) (*Nate Cook*)	(Jieun Kim, *Yoonseo Kim*)	groupings that have been overlooked? (*Tove Larsson*, Doug Biber, Tony Berber Sardinha)
5:15-6:45pm		Plenary Talk 4: Leveraging "AI", open science, and analytic expertise in corpus linguistic research (*Kristopher Kyle*) (introduction: *Ute Roemer-Barron*)			
6:45-7pm		Closing Remarks			

List of Poster Presentations

#	Title	Authors
2	Contrastive analysis of mood and modality in English and Naija	Abiola Iyiade
9	A corpus-based contrastive study on the behavioral profile of adverbs of degree in Mainland and Taiwan Chinese	Yan Xiao
16	Overcoming database and software challenges in corpus linguistics research in developing countries: a case study of Africa	Solange Swiri Tumasang
20	Temporal fluency and lexical diversity: a comparison of their predictive power on the Spanish OPIc	Alan Brown, Greg Thompson, Troy Cox, Earl Brown
23	Noun phrase development across registers in the first two years of Chinese learning	Yilei Li
28	Can AI sound like a teacher? Corpus evidence from human and simulated classrooms	Marilisa Shimazumi, Tony Berber Sardinha
29	Rewriting immigration: discursive shifts in AI-generated news	Tony Berber Sardinha, Marilisa Shimazumi
36	A negotiated communion in parliamentary debates: concerns and representations of hospice-related objects in Hansard	Yan Cheng
37	Lexical appropriateness of the Korean College Scholastic Ability Test (CSAT) English: a comparative analysis with English Corpora	Nena Choi
45	Zooming in and out on the learner lexicon: a CDST approach to L2 Turkish vocabulary development	Yigit Savuran, Stefanie Wulff
53	A corpus analysis of semantic drift and pejoration in English	Chad Hammock
57	LC-meta: a core metadata schema for L2 data documentation	Jennifer-Carmen Frey, Larissa Goulart da Silva, Alexander König, Hubert Naets, Egon Stemle, Magali Paquot
59	Developing a corpus of reading errors in English early child language	Madison Rose, Michael Bennie, Valeria Pagliai, Walter Leite, Zoey Liu
61	Lexical multidimensional analysis of art discourse: capturing human experience in the language of Sally Mann's photography	Yara Maria De Toledo Dias Romeiro
70	Comparing computer-based and paper-based DDL for vocabulary learning	Dilay Candan, Senem Yıldız Ersoy, Ute Römer-Barron
78	Enlivening script to stage: toward a multimodal corpus-based approach to theatre translation studies with Dou E. Yuan as an example	Shi Li
79	Changes in referential production among Japanese-English bilingual returnee children: a five-year longitudinal study	Jason Rothman, Maki Kubota, Stefanie Wulff, Vicky Chondrogianni
80	Automatic analysis of second language learners' development of verb-argument constructions in a longitudinal writing corpus	Soyeon Sim
81	Validating the Phraseological Complexity Analyzer (PaCa): a digital tool for assessing phraseological diversity and sophistication	Shuyuan Tu, Daniel H. Dixon

87	Thematic progression anomalies in Japanese university students' argumentative essays: a functional linguistic perspective	Tsukuru Kamiyama
88	Lexical organization in learner production: the role of frequency, context, and semantics	Anton Vegel
91	Pronoun use in compliments: capturing cultural variation in human experience	Jini Jung
92	A computational diachronic analysis of gen-Z mental health discourse: a large-scale reddit corpus study from pre- to post- COVID	Felix Mao
93	Do keywords tell a different story? A Comparison of key feature and keyword analyses in research article introductions	Nergis Danis
95	Linguistic differences in humor: a feature-based comparison between human and AI-generated jokes	Freya Pan
96	Mapping narratives of resilience in the heat–health–climate nexus: insights from a corpus linguistics approach	Ersilia Incelli
98	Context valence as a tool for categorising semantic prosody: investigating register-sensitivity and the impact of polysemy	Mathias Russnes
99	The sociolinguistic impact of divergence in human experience: variation and change in Laurentian French	Hélène Blondeau, Raymond Mougeon, Mireille Tremblay
108	Leveraging masked language models to measure association strength in contiguous and dependency bigrams	Hakyung Sung, Kristopher Kyle
114	What are hybrids? A multidimensional analysis of hybrid texts on the French and Swedish web	Saara Hellström, Erik Henriksson, Veronika Laippala
120	Linguistic creativity and spontaneity: an investigation of hypothetical *if*-clauses	Yen-Po Chen, Siaw-Fong Chung
121	Clustering embeddings from register classifiers reveals fine-grained structure within web registers	Erik Henriksson, Tuomas Lundberg, Antti Kanner, Veronika Laippala
123	Statutory interpretation of a verb using prototype-by-component analysis	Alyssa Lowry
124	The role of register range in explaining lexical decision times	Daniel Keller, Earl Brown, Brett Hashimoto
126	Investigating the use of acoustic cues for addressee inference in machine-directed speech	Cassidy Henry, Alayo Tripp, William Idsardi
129	Corpus-informed prompt design for LLM-mediated support for researchers with limited proficiency in Academic English	Tony Berber Sardinha, Ana Boconry, Deise Dutra, Walcir Cardoso, Anderson Ávila, Marilisa Shimazumi, Vivian Lameira, Juliana Almeida, Luciana Aguiar de Oliveira, Gabriela Escobar
131	Affective language and fake news in Brazilian Portuguese	Camila Lívio, Chad Howe
133	Mapping words to functional clusters with prosodic profiles	Ryan Ka Yau Lai, Lu Liu, Haoran Yan, John DuBois
137	Applying Corpus-Assisted Discourse Studies (CADS) to examine representations of agriculture in *The Catholic Worker*	Shaya Kraut
138	Implicit connectives are also eRST signals!	Lin Ai, Amir Zeldes

140	Lexical markers of climate discourse in the Polish opinion press: a corpus-based and discourse-historical approach	Dagmara Mateja
143	NLP tools for constructing a spoken corpus of endangered language varieties: a case study of the ELIC Corpus	Massimo Daul, Austin Jones, John Hale, Margaret Renwick, Zvjezdana Vrzić, Keith Langston
146	From “proposing” to “arguing”: academic writing in an English major program	Marine Matte, Simone Sarmento
148	Testing a TxTLx approach to variation in dissertation writing	Febriana Grundy
149	Comparing reporting verb use across L2 student writing and applied linguistics articles: a replication study	Duong Nguyen, Men Truong
156	CRetor: an annotated corpus of rhetorical strategies in Mandarin counter speech	Xiaoyu Chen, Chenfeng Su, Michael Bennie
157	Register-based trends in the grammatical complexity of student writing	Bethany Gray, Duong Nguyen, Febriana Lestari, Kimberly Becker
159	Framing fertilization: a corpus analysis of gamete-centered language and metaphor in Japanese	Lauren Polak, Kaori Idemaru, Cindi SturtzSreetharan
161	College entrance written exam analysis from a multidimensional perspective: a corpus linguistics approach	Juliana Barreto
162	From syntax to discourse: LLM-assisted annotation of code-switching in typologically diverse language pairs	Olga Kellert
163	Not just “may” and “might”: mapping multi-word hedges in research articles	Wesley Acorinti, Alexander Holmberg

Plenary Speakers

Kenji Sagae **(University of California at Davis)**

COWS-L2H: A Corpus of L2 and Heritage Spanish at the Intersection of Computational Linguistics and Language Instruction

Data-driven study of language learner behavior requires specialized corpora. Such datasets are also increasingly valuable in the development of educational technology applications for language learners. Many university language courses generate a steady stream of written language data, but efforts to leverage courses in the creation of language resources are relatively rare. In this talk, I will describe COWS-L2H, a corpus of L2 and heritage learner Spanish that is the result of years of collection of essays written by students who volunteered about 3,500 writing samples while taking Spanish courses at various levels at UC Davis, ranging from beginning to advanced, in addition to courses aimed at heritage speakers. In addition to nearly 1.4 million words, the dataset includes anonymized metadata, including first language, course level, gender, and age. A key aspect of the corpus is the inclusion of a substantial amount of longitudinal data from students who contributed essays over multiple academic terms while enrolled in a course sequence.

I will discuss the methodological challenges and considerations associated with data collection and annotation, as well as recent research results obtained using the corpus. The structure of the curriculum and the scheduling of essay prompts that cycle over different academic terms allow for both a longitudinal view of language learning and controlled comparisons across prompts. I will also discuss the use of COWS-L2H as a testbed for computational systems, including its use in research on automatic grammatical error correction, and in a shared task where teams from different universities developed various types of computational models for tracking language development using the same corpora. Finally, at a time when language datasets in specialized domains, genres and tasks are highly valued for their use in technology based on neural language models, COWS-L2H illustrates how universities may leverage the intersection of research and teaching to benefit from data created within themselves to go beyond generic commercial offerings and increase their agency in the development of the educational language technology used by their students.

Sali A. Tagliamonte **(University of Toronto)**

Community Corpora for Linguistic Science: Studies from a Project about History, Culture and People

In the context of global communication networks, declining traditional varieties and dramatic cultural shifts, what kind of corpora can inform the advancement of linguistic science? In this presentation, I synthesize insights from a large-scale documentation project in Ontario Canada, the Ontario Dialects Project (Tagliamonte 2014; Tagliamonte 2018). The data comprises vernacular speech from fieldwork in 21 communities across the province, amounting to over 14 million words from 1525 individuals born from the late 1879-2011. The data come from 21 different locales, including a range of urban-non-urban locations, dense vs. loose social networks, varying population sizes, ethnic compositions, founding populations, etc. covering many of the community types described in Trudgill (2011). The project is part of a larger research program which aims to advance the knowledge base of language variation and change in time and space as well as to provide front-facing secular publicity on the importance of language to history, culture and identity.

The sheer size of the Ontario Dialects Project holdings permits study of a broad range of linguistic phenomena at different levels of language, from words and expressions to grammatical systems; from frequent features to rare constructions, many of which are moribund in mainstream varieties. Using a selection of case studies from recent research, I demonstrate how the findings arising from studying vernacular language in a large corpus with regional and social nuances provide insight into key contemporary questions in linguistics, including diachronic processes, linguistic innovation, language contact, place and identity. Taken together, the discoveries emerging from the Ontario Dialect Project offer a comprehensive perspective on synchronic patterns of language that can be used to inform future research as global varieties continue to emerge— geographically situated and not — offering a vital documentation of the evolution of human language in its ever-changing networks of social interaction.

Selected References

Trudgill, P. J. (2011). *Sociolinguistic typology: Social determinants of linguistic complexity.* Oxford: Oxford University Press.

Tagliamonte, S. A. (2014). System and society in the evolution of change: The view from Canada. In Green, E. & Meyer, C. (Eds.), *Variability in Current World Englishes.* Berlin and New York: Mouton de Gruyter. 199-238.

Tagliamonte, S. A. (2018-2024). Language change and social change in the early 21st century: Canadian English 2002 to 2020. #435-2019-0053. Toronto, Canada: Social Sciences and Humanities Research Council of Canada.

Paul Baker **(Lancaster University)**

A picture is worth 1000 words? A multimodal investigation of an image-tagged corpus

Until recently, full-scale multimodal analysis of news corpora has been difficult to realise, due to the scale of annotating thousands of images along with finding a way to integrate such images into a linguistic analysis. Many analysts have either ignored images or carried out a separate, small-scale analysis of images. However, advances in AI-based image tagging offer the potential to achieve a full multimodal analysis. This talk describes the multimodal analysis of a 1.5 million word corpus consisting of a year's worth of UK national press news articles about Islam and Muslims, published between December 2022 and November 2023 (Baker et al 2025). The corpus also contains 8,546 image files which were automatically tagged using Google's Vertex AI. I first describe how the corpus was created and tagged, and how the image tagging was evaluated and improved upon to reach approximately 90% accuracy. The corpus contains articles taken from eight newspapers and the central research question involved identifying features that were distinct to each newspaper. In order to answer this question I outline how analysis was carried out on three levels: a) written text only b) images only c) interactions between written text and images. For example, investigation of keyness indicated a mixture of content keywords like *Ramadan*, *jihad*, *BBC*, *Palestine* and *Koran*, along with grammatical keywords like *her* and *you*. On the other hand, key image tags included some which appeared to reference social actors like Tie, Speech, Ear and FashionModel, while others appeared less obvious: Gas, Rectangle, Darkness, Landmark. Keyness analysis was supplemented with consideration of collocates e.g. what words are likely to occur near or next to a particular image if it has been assigned a certain tag. Or what kinds of images were likely to occur in the vicinity of a particular keyword?

I compare the three levels of analysis together with the aim of investigating the extent to which consideration of images provides 'added value' by identifying new insights into press discourse around Muslims and Islam or simply confirms the analysis of written text. I demonstrate the affordances of the three approaches, providing a critical evaluation of Vertex AI's capabilities, the value of automatically-assigned image tags, the abilities of popular corpus software to work with visually tagged corpora and the potentialities for further image-tagged corpus research.

References

Baker, P. Scheumk, H and Qian, Y. (2025) *Automatic Image Tagging for Corpus Linguistics: A multimodal study of news representations of Islam*. Cambridge: Cambridge University Press.

Kristopher Kyle **(University of Oregon)**

Leveraging "AI", open science, and analytic expertise in corpus linguistic research

Corpus linguistics is broadly concerned with describing language use based on representative samples of spoken, written, and/or signed texts. While there are notable exceptions (e.g., Hunston, 2022; Sinclair, 1991), most corpus linguistic research leverages linguistic annotation of some sort (e.g., part of speech tags, syntactic dependency relations, and/or other information). As the growth of the internet has made the collection of extremely large corpora feasible for most researchers, linguistic annotation has increasingly been automated (e.g., Davies, 2009; Schäfer, 2015; Wenzek et al., 2020). Corpus linguists have leveraged standard NLP tools such as POS taggers and syntactic parsers for a variety of tasks such as homograph disambiguation (e.g., Jarvis & Hashimoto, 2021; Kyle et al., 2018), analysis of collocations (Granger & Bestgen, 2014; Kyle & Eguchi, 2020; Paquot, 2019), and the identification of particular grammatical features (e.g., Biber, 1988; Kyle & Crossley, 2017, 2018; Lu, 2011) among many others.

Nonetheless, the utility of taggers and parsers in linguistic analysis is not without limit. Linguistic features that are formally ambiguous (lexically and with regard to grammatical category and syntactic structure) such as argument structure constructions (ASCs; Goldberg, 1995), many discourse features (e.g., engagement strategies; Martin & White, 2005), etc., cannot be disambiguated using such tools. Accordingly, the analysis of many linguistic features of interest has required either manual analysis (e.g., of a manageably sized random sample) or the development of novel annotation tools. However, annotation tool development has been relatively limited in the field of corpus linguistics for at least three reasons. First, earlier machine learning models were limited in their ability to incorporate sufficient contextual information which constrained the range of linguistic features that could be accurately annotated. Second the development of accurate automated annotation models typically required large amounts of manually-annotated training data. Third, the development of machine learning algorithms has not typically been part of a corpus linguist's training.

The advent of "AI" (e.g., pre-trained language models [PLMs] and large language models [LLMs]) and the growth of the open science movement, however, have largely mitigated the barriers to the development of new annotation tools. In this talk, I will discuss how newly developed machine learning tools ("AI"), in concert with affordances of the open science movement, have opened the door to the automated analysis of a wide range of linguistic features and how corpus linguists are uniquely suited to pursue this work. I will also present two concrete examples of related work and empirically demonstrate that reasonably accurate annotation models of formally ambiguous linguistic features can be developed with a small amount of manually annotated training data. Finally, I will present a framework for collaborative annotation projects, introduce an ongoing annotation effort, and invite wide collaboration from the field of corpus linguistics.

Presentation Abstracts

#12

Xenophobic representations targeting China on “X" during the COVID-19 pandemic

Cicero Soares da Silva

In this paper, we look at malicious representations of China on Twitter occurring in the context of the COVID-19 pandemic. In order to capture these detrimental representations, we scraped a corpus of ca. 100K tweets in Brazilian Portuguese containing ten highly xenophobic hashtags, which were used by right-wing followers to discredit China and spread hatred. The multimodal method followed in this study consisted of a combination of Lexical Multidimensional Analysis (LMDA; Berber Sardinha & Fitzsimmons-Doolan, 2024) and Visual Multidimensional Analysis (VMDA; Berber Sardinha et al., 2023). The LMDA used lexical units to detect traces of discourses across the texts, whereas the VMDA applied computer vision AI techniques to annotate the images posted along with the twitter messages. Two sets of dimensions were obtained (i.e. verbal and a visual). Six verbal dimensions were identified, the first two being: (1) Pandemic manipulation vs Pro-president hashtags: this contrasts the alleged manipulation involving the protection of corrupt officials and misleading pandemic data with the use of hashtags to ridicule China, support presidential policies, and target political adversaries. (2) System and media rejection hashtags vs Anti-China normalization: This captures the use of hashtags to reject the political system and its alleged ties with China, versus efforts to normalize anti-Chinese sentiments under the guise of common sense and cultural distortion. In turn, five visual dimensions were determined, the first two being: (1) China scam denouncement vs. Brazil Elite Accusations: This captures the contrasting positions of denouncing scams from China and accusing Brazilian intellectuals of siding with Chinese interests; (2) Weak local government pandemic response vs. Brave pandemic leadership: This contrasts the views on local governments' responses to the pandemic, either as weak and closure-promoting or brave and against closures seen as benefiting Chinese interests. All the dimensions will be discussed and illustrated in the paper presentation.

References

Berber Sardinha, T., & Fitzsimmons-Doolan, S. (2025.). Lexical Multidimensional Analysis. Cambridge University Press.

Berber Sardinha, T. (2024). Exploring multimodal corpora in the classroom from a multidimensional perspective In P. Crosthwaite (Ed.), Corpora for Language Learning: Bridging the Research-Practice Divide (pp. 25-36). Abingdon: Routledge."

#17

Variability within stability: a novel key keyword multidimensional analysis of the diachronic coverage change in China Daily's international media reporting on China (2017-2024)

Chenghui Wu

While diachronic corpus linguistics has excelled at identifying linguistic variability, the corresponding concept of stability—the persistent lexical architecture that forms the bedrock for change—remains a significant methodological challenge. This paper addresses this gap by proposing and applying a novel framework, Key Keyword Multidimensional Analysis (KK-MDA), to investigate how the representation of human experience varies within a stable discursive context. We demonstrate this approach through a large-scale analysis of China Daily, tracking how the newspaper strategically shifted its portrayal of experience from the collective to the individual over an eight-year period (2017–2024).

Our primary contribution is methodological. The KK-MDA framework operationalizes the concept of 'variability within stability' through a bespoke, multi-stage procedure performed on a self-built diachronic corpus of 800 news reports (871,325 tokens). The procedure begins by identifying a comprehensive pool of candidate keywords. Adapting the Keyword Co-occurrence Analysis (KCA) methodology, we conduct a series of rolling year-on-year comparisons, which effectively capture words that become newly salient or significantly increase in use, thus signaling linguistic change. The core innovation lies in the subsequent stage: this large candidate pool is subjected to a rigorous stability filter. We employ the Deviation of Proportions (DP) measure, a sophisticated metric of distributional evenness, to quantify how consistently each keyword is used across the eight annual sub-corpora relative to their size. Based on a data-driven optimization process, we isolate a final set of 81 exceptionally stable keywords that form the persistent lexical core of the discourse. Finally, a Principal Component Analysis (PCA) is performed on a text-by-keyword matrix of these 81 stable key keywords to uncover their underlying co-occurrence patterns.

This meso-level analysis reveals four stable, underlying discursive dimensions that function as rhetorical frames: (1) National Strategy & Development vs. Individual & Daily Life; (2) Regional & Domestic Affairs vs. Global & Identity Narratives; (3) Discourse of Retrospective Achievement; and (4) Discourse of Prospective Initiatives. These dimensions are distinct from both the micro-level stylistic features typical of traditional corpus-based Multidimensional Analysis and the macro-level topics often identified by keyword analysis. They represent the rhetorical angles through which any topic can be framed.

The diachronic analysis of these dimension scores reveals a profound variation in the framing of human experience. The most dramatic finding is a strategic pivot along Dimension 1. The discourse peaked with a strong focus on macro-level, state-centric narratives in 2018, before plummeting to a pronounced emphasis on micro-level, personal storytelling in 2022. This demonstrates that China Daily's coverage change was not merely topical but a fundamental rhetorical reorientation. By first foregrounding stability, our KK-MDA framework effectively captures how core institutional concepts are dynamically re-framed over time, revealing a deliberate strategy to vary the representation of human experience to engage a global audience. This study offers a new methodological pathway for diachronic discourse analysis and provides substantive insights into the strategic evolution of international state media.

#18

From shielding lives to growing futures:
a corpus-based study of multimodal metaphors in Korean insurance posters

Ebru Turker

While corpus-based research has traditionally focused on lexical and grammatical variation, corpora that integrate both image and text remain relatively infrequent, and to date, there is no dedicated image–text corpus for Korean. This study addresses that gap by constructing and analyzing a specialized multimodal corpus of Korean insurance company posters, integrating verbal and visual data to investigate how metaphor operates across modes in commercial discourse.

The dataset, compiled from the publicly accessible portion of the KOBACO Advertising Museum Digital Archive, includes Korean insurance company posters produced over the past decade. These materials—distributed widely in both physical and digital public spaces—employ emotionally charged imagery and persuasive textual elements to frame insurance as protection, growth, shelter, and care. Corpus construction involved targeted keyword searches (보험 'insurance,' 생명보험 'life insurance,' 자동차보험 'auto insurance'), with each item cataloged for year, advertiser, campaign title, and medium.

Both verbal and visual elements were systematically annotated using a coding scheme based on Forceville's (2006, 2009) multimodal metaphor framework. Verbal text was transcribed, segmented, and tagged for metaphor-related lexical patterns; visual content was described and coded for metaphor source–target mappings, multimodal alignment (reinforcing vs. complementary), and cultural resonance. The annotation also recorded emotional tone (e.g., warm, protective, aspirational) and sociocultural references (e.g., collectivism, filial piety, intergenerational solidarity).

Preliminary findings indicate recurrent conceptual metaphors such as LIFE IS A JOURNEY, SECURITY IS A SHIELD, and FUTURE IS A GROWING PLANT. These metaphors appear verbally in expressions like "prepare for tomorrow" and "grow your future," and visually in seedlings nurtured by hands, families enclosed in domes, and bridges or ladders as pathways to the future. Frequency and co-occurrence analyses reveal notable patterns—for example, growth metaphors frequently co-occur with family imagery, while shield metaphors tend to accompany verbal appeals to safety and trust. These patterns illustrate how Korean insurance advertising strategically integrates modes to reinforce culturally resonant meanings.

By foregrounding the interaction between verbal and visual modalities in an empirically derived corpus, this research demonstrates how multimodal metaphor functions both as a rhetorical device and as a meaning-making resource in Korean commercial discourse. The methodological approach—building a corpus from public digital archives, systematically annotating across modes, and combining qualitative and quantitative analysis—offers a replicable model for corpus-based studies of multimodal discourse. This work not only expands the scope of corpus linguistics into image–text data but also contributes to cross-cultural advertising research by showing how metaphors align institutional messaging with culturally salient values in Korean society.

#21

Interdisciplinary applied corpus linguistics:
a case study of a promising partnership and some questions for the field

Shannon Fitzsimmons-Doolan, Jennifer Beseres Pollack

This presentation will focus on the promise of interdisciplinary partnerships between corpus linguists and scholars from a range of other disciplines to generate meaningful applied scholarship. It begins by introducing the field of interdisciplinarity, its tenets, and then moves on to highlighting key considerations of effective interdisciplinary collaboration generated by the field (Aboelela et al., 2017; Repko et al., 2020). Next, a case study—focusing on logistics—of the presenters' collaborative research agenda integrating applied corpus linguistic and marine ecology serves as a central example.

The three interdisciplinary applied linguistics studies in the focal collaboration (Authors 1, 2023; Authors 2, 2025; Authors 3, in preparation) investigate texts related to oyster resource management and restoration in the Gulf of Mexico. For each study, the presenters developed research questions through in-depth discussion of the relevance of the texts, hypotheses informed by theory from both disciplines, and the analytical affordances of various corpus linguistic techniques. The results inform communication strategies and regional management efforts for coastal and estuarine ecosystems. Lessons learned from the collaboration include paying keen attention to project audience at the outset, engaging in explicit communication about many usually taken-for-granted decision points, and being proactive about publishing venues. These and other lessons will be elaborated upon in the presentation.

Based on the information presented, this session will conclude by asking questions to the field. What other interdisciplinary applied corpus linguistics projects are being conducted? How are their findings being disseminated? What information is needed to support more interdisciplinary applied corpus linguistics scholarship?

#22

Teaching algebra online:
patterns of teacher talk and their relationship to student outcomes

Zhihui Fang, Rui Guo, Zifeng Liu, Wanli Xing

Mathematics education faces a significant challenge: many students find the subject daunting due to its complex interplay of specialized language, symbols, and visual representations (Fang, 2024). This complexity makes teacher explanation paramount, as students' construction of mathematical knowledge is heavily dependent on instructors' oral discourse (Moschkovich, 2021; Schleppegrell, 2007). This dependence is amplified in online learning environments, where video lectures are the primary medium for instruction. Consequently, understanding the patterns and efficacy of teacher talk in these settings is crucial.

Grounded in Systemic Functional Linguistics (SFL), a sociosemiotic theory that views language as a dynamic resource for making meaning and provides tools for analyzing language use in context (Halliday & Matthiessen, 2014), our study investigated how teachers explain algebraic concepts and ideas in video lectures and how their discursive patterns correlate with student outcomes. Three research questions were addressed: (1) How do teachers explain mathematical concepts? (2) Are there significant differences in their explanation styles? (3) Is there a correlation between teachers' talk patterns and student learning?

Data were sourced from Math Nation (an online platform for secondary school mathematics) and consisted of 125 video lectures on 25 algebra topics (e.g., polynomial expressions, quadratic expressions, exponential functions, algebraic expressions), delivered by five instructors with varying academic background (e.g., statistics, sports, mathematics, education, business administration) and teaching experience (from 0 to 20 years). Each video lesson lasted 10 to 40 minutes. These lessons were analyzed for their language patterns using SFL tools previously piloted in Xing et al (2025). Specifically, logical analysis focused on the use of elaboration, extension, and enhancement to assess how logical-semantic links were built across utterances within each lesson. Grammatical analysis focused on verb types (e.g., material, mental, relational), participant categories (mathematical entities, students/teacher, other), and lexical density (e.g., technical words, complex noun phrases) to assess how mathematical concepts, ideas, and relationships were represented in each lesson.

MANOVA was used to determine potential differences among teachers in their language use. Correlation analyses (Pearson's or Spearman's) were used to examine the relationships between teachers' language use and two key student outcomes: engagement (measured by in-video dropout rates) and learning (measured by post-lesson assessment accuracy).

Data analysis is ongoing and is expected to be completed by the end of October, 2025. Preliminary findings shed light on how contrasting teacher expertise levels manifest in their linguistic mediation of technical content within online mathematics instruction. Specifically, compared to novice teachers, more experienced teachers used more instances of elaboration and enhancement and less complex but more engaging grammatical resources when explaining algebraic concepts and ideas. However, their lessons also tended to be last longer, leading to lower student engagement. Despite these differences, there were marginal differences in student learning outcome. These findings unravel the complex relationship among discourse patterns, engagement, and learning outcome.

Our study provides valuable insights into how teachers present algebraic concepts and ideas in ways that both simplify complexity and enhance explanatory clarity and depth. These insights should benefit those seeking to strengthen communication effectiveness in online mathematics instruction.

#26

From rates of occurrence to prototypicality: mapping the distinctive features of conversation

Elizabeth Hanks

In most corpus research, 'conversation' is presented as largely homogeneous and clear-cut. However, Hanks (2023) suggests that conceptualizations about conversation are largely inconsistent regarding purpose (social and/or task-oriented), modality (spoken and/or written), and production (planned or unplanned). This indicates that the register of conversation may be broader and more heterogenous than commonly assumed. Building on this work, the present study draws on prototype theory to conceptualize texts on a continuum from highly prototypical to highly peripheral to the register of conversation. This study examines which linguistic features are most prototypical of the register of conversation in American English by (a) extracting frequencies of linguistic features in a corpus (e.g., phrasal complexity, pauses), (b) collecting perceptual data on the extent to which interactions are considered prototypical of conversation, and (c) analyzing connections between the two.

Qualitative and quantitative analyses were conducted to determine the range of situational characteristics that conversation may exhibit. Based on this analysis, the Corpus of Contexts Across Registers and Texts (CCART) was compiled to represent the situational characteristics of conversation as inclusively as possible. CCART contains 153,526 words across 1,380 discourse units (DUs; units of language consisting of uniform communicative goals; Egbert et al., 2021) that were manually identified from 18 existing corpora and databases.

Rates of occurrence for 52 linguistic features were calculated through the Biber Tagger (e.g., Biber, 1988) and additional Python and R scripts. Each DU was then rated by 2 to 9 participants from a total pool of 898 survey participants. Drawing on methods used in prototype theory (e.g., Rosch et al., 1976), participants rated DUs according to the extent to which they are prototypical examples of conversation on a seven-point scale. Ratings for each discourse unit were averaged to result in a prototypicality score that reflects the extent to which each DU is prototypical of conversation.

Continuous key feature analyses (Egbert et al., 2026) that utilized both correlation and multiple regression were conducted to determine which linguistic features of conversation are most and least distinctive of prototypical conversation texts. Among the linguistic features, frequent turns, faster speech rate, and more balanced participant contributions emerged as some of those most distinctive of prototypical conversation texts. Notably, some of the linguistic features that are shown to be most frequent in conversation are not highly distinctive of conversation. For example, adverbs are one of the most frequent linguistic features in conversation, yet the results from perceptual data indicate that adverbs are not distinctive of prototypical conversation. It is possible that frequent yet unimportant features such as adverbs have low cue validity, resulting in decreased cultural salience (Murphy, 2002).

The results of this study highlight the heterogeneity of conversation, and the methods used demonstrate how cultural insights from members of the general public can be used to understand register as a culturally-recognizable construct on the level of the register, the text, and the feature. These findings can promote consistent dataset descriptions and enhance comparability of both results and corpora across studies.

#32

Text-level grammatical complexity revisited:
Are there additional groupings that have been overlooked?

Tove Larsson, Doug Biber, Tony Berber Sardinha

Grammatical complexity features, which add optional structural elements to 'simple' clauses or phrases, have been studied extensively in recent years (see Biber et al., 2025, for an overview). These features range from dependent phrases functioning as noun phrase constituents (e.g., attributive adjectives, pre-modifying nouns) to finite dependent clauses functioning as clause constituents (e.g., finite adverbial clauses, verb + that-complement clauses). It has been established that these features do not occur independently of each other across texts and registers. Rather, groupings of complexity features systematically co-vary across texts. For example, if a text uses frequent finite adverbial clauses, that same text is likely to use frequent verb + wh-complement clauses (e.g., Biber et al., 2022).

However, the exact nature of these groupings has yet to be established. Biber, Larsson, & Hancock (2024a, b) hypothesized that groupings based on shared structure and syntactic function would correspond to sets of features that tended to co-occur in texts. Those studies found evidence to support the groupings of prototypical phrasal and clausal complexity features, but a majority of the remaining complexity features were left unaccounted for with this method. That is, across texts, there are only weak covariance patterns among many complexity features. These findings suggest that there are still unknowns with regard to whether there are other, so-far undetected groupings of complexity features at the text level.

The present study builds directly on the findings of prior studies, with the goal of identifying subsets of grammatical complexity features that covary at the text level. To this end, we used a multi-register corpus (~7.3 million words) that was carefully compiled to cover a broad range of contexts and situational characteristics in both the written and spoken mode. We used Multi-dimensional Analysis (Biber, 1988) to identify subsets of features that covary across texts, interpreted as underlying dimensions of variation. In the first place, the analysis shows that 10 of the 25 complexity features included in the study failed to co-vary in systematic ways with other complexity features, confirming the previous conclusion that many complexity features are distributed in their own unique ways. Beyond that, the results show that three dimensions capture 45 percent of the variance among the remaining 15 features. We compare the composition of those dimensions to the groupings identified in previous research, discussing similarities and differences relative to the particular goals of each analytical approach.

#33

Exploring the humanlikeness of AI-generated texts through register alignment: a corpus-based study of ChatGPT

Yağmur Demir

The rise of AI tools like ChatGPT has transformed language pedagogy and assessment. While AI is believed to produce human-like language, concerns remain about the quality of language in AI-generated texts used for educational purposes. Despite growing research, our understanding of the humanlikeness of AI-generated texts remains limited, particularly from the perspective of register which is a key feature of human language. Register serves as an important factor in understanding and assessing the humanlikeness of AI-generated texts (Berber Sardinha, 2024). Therefore, to determine whether AI-generated texts are truly human-like, we must assess their ability to reflect the register-specific linguistic characteristics of human language.

The current study investigates the register alignment of the AI tool ChatGPT through a comprehensive corpus linguistic analysis by addressing the following research question: To what extent do ChatGPT-generated texts align with the linguistic variation found in human-authored texts across different target registers (research articles, university textbooks, and Wikipedia) and disciplines (biology and political science)?

The corpus consists of 600 texts: 300 human-authored texts and 300 ChatGPT-generated texts with 100 texts per register (50 texts per discipline). Register alignment is examined statistically using an additive multidimensional analysis (MD) of Egbert's (2014) study, as well as a semantic key feature (KF) analysis. Effect size benchmarks determined specifically for this study (i.e., Cohen's d) are used to evaluate the magnitude of both similarities and differences in register alignment.

Findings of MD analysis indicate that ChatGPT-generated texts have a harder time mimicking the university textbook register, and particularly, the political science discipline across all registers. These texts tend to use denser and more technical language and exhibit less author-stance. Effect sizes show a strong misalignment between ChatGPT-generated and human-written texts in key discourse functions such as defining and evaluating new concepts.

Findings from the semantic KF analysis show that ChatGPT exhibits a similar semantic profile across registers and disciplines. Same categories repeatedly emerge as key features in ChatGPT subcorpora. For example, ChatGPT consistently relies more heavily on abstract and cognitive nouns, facilitation/causation verbs, characteristic, topical, and evaluative adjectives, as well as approximator adverbs than human authors.

Together, these findings suggest that AI-generated texts diverge from the linguistic variation found in human-authored texts. As a result, they do not exhibit a linguistic profile that aligns with the situational characteristics of target registers (i.e., functional appropriateness). In other words, ChatGPT tends to produce a "one-size-fits-all" academic writing style that lacks the nuanced linguistic characteristics of human writing.

These results highlight the need for further quantitative linguistic analyses of AI-generated language, especially when AI is used for educational content creation. Moreover, they carry important implications various stakeholders.

#40

LLMs and ideological priming: a case study of immigration discourse

Tony Berber Sardinha, Anderson Avila, Maria Claudia Nunes Delfino, Marilisa Shimazumi, Deise Prina Dutra, Paula Tavares Pinto, Ana Bocorny, Carlos Kauffmann, Patricia Bértoli, Mirella Whiteman, Marcos Roberto de Oliveira, Rogerio Yamada, Leandro Tessler

Large Language Models (LLMs) have been used to generate texts that end up in circulation in society, thereby influencing people's worldviews or ideologies. Given that LLMs can produce ideological discourses, it is reasonable to ask to what extent LLMs can be primed to reverse or reinforce ideological positions. To throw light onto this issue, this study investigates how GPT 4o re/produces discourse when rewriting news articles about immigration under three conditions: convergent (preserving the original ideological stance), divergent (reversing the stance), and internal knowledge (no ideological priming). The human-written source texts were selected from US and UK news media, and pre-categorized as either right- or left-wing. Each text was rewritten by the LLM under all three conditions, yielding a corpus of 9600 texts (ca. 3,500,000 words). To describe the underlying ideologies in the texts, we employed Lexical Multi-Dimensional Analysis (LMDA; Berber Sardinha & Fitzsimmons-Doolan, 2025), which enables the description of ideological discourses in corpora through lexical pattern analysis. Two separate LMDAs were conducted. The first identified the discourses underlying the human-authored texts, yielding five dimensions: (1) nationalist and security-oriented discourse versus humanitarian advocacy and integration; (2) discourse of sovereign enforcement versus humanitarian inclusion; (3) post-colonial justice and structural reparation versus securitized and populist defense; (4) migration trauma and moral accountability versus regulatory and fiscal liability; and (5) historical oppression and migrant vulnerability versus law-and-order criminalization. These were built into prompts that queried the AI to generate texts based on each writing condition. The output from the LLM was collected, forming a second corpus. This AI corpus was aggregated with the human texts, which was also submitted to an LMDA to detect the corresponding underlying discourses. Seven dimensions emerged, contrasting discourses of human narration and lived experience versus institutional and sovereign authority (Dim. 1), moral imperative and victimhood versus dispute and polarization (Dim. 2), cultural enrichment and postcolonial justice versus illegality, securitization, and moral panic (Dim. 3), regulatory nationalism versus emotionally charged advocacy (Dim. 4), assertive leadership and cultural clash versus civil oversight and condemnation of state handling (Dim. 5), law and order and executive action (Dim. 6), and humanitarian crisis and moral demand (Dim. 7). The texts were scored on each dimension, enabling the comparison of ideological positions (left-wing vs right-wing), task conditions, and source (human vs AI). These comparisons showed the following. By source, AI-generated texts were more marked across all dimensions than human-authored ones, with the exception of Dimension 1, which depended on informal, conversational language, showing that the AI representation of news discourse defaults to a formal style. By ideology, the AI proved flexible enough to reinforce both left- and right-wing discourses. By task type, the rewrite tasks strengthened the ideological discourses relative to the internal knowledge baseline, suggesting the AI can be steered to exacerbate its ideological stance if so prompted. Finally, the study provides evidence that LMDA dimensions can be effective devices for AI ideological priming.

Reference

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

#41

Do first-year composition assignments align with the linguistic and situational demands of disciplinary writing?

Larissa Goulart, Lizzy Hanks

Composition courses are designed to help undergraduates transition from general to academic writing (Aull, 2015). These courses often serve as students' first exposure to the conventions of academic discourse and the types of assignments required across the curriculum. However, research has long questioned whether such courses effectively develop transferable skills for disciplinary writing (Bridgeman & Carlson, 1984; Leki & Carson, 1997). More recent studies continue to investigate this issue, reporting mismatches between the linguistic characteristics of composition and disciplinary writing (Staples & JoEtta, 2022; Goulart & Hanks, 2024). Building on this work, the present study compares the situational and linguistic characteristics of first-year composition writing with those of three disciplinary groups: life and physical sciences, social sciences, and arts and humanities. The composition corpus includes 90 texts written for two first-year courses, while the disciplinary corpus comprises 432 texts from MICUSP, divided across the three disciplinary groups. For the situational analysis, texts were coded by three annotators for communicative purposes (e.g., explaining, comparing, narrating; Goulart et al., 2022) as well as layout features (e.g., abstracts, visuals, references, citations). Descriptive statistics were used to compare the extent to which texts exhibit these situational features across first-year composition and the three disciplinary groups. For the linguistic analysis, we conducted an Additive Multidimensional Analysis (MDA) based on Biber's (1988) dimensions: (1) Involved vs. Informational Production, (2) Narrative vs. Non-Narrative, (3) Explicit vs. Situation-Dependent, (4) Overt Expression of Persuasion, and (5) Abstract. This approach allows for the analysis of co-occurring linguistic features within the present dataset as well as enables cross-study comparisons with other applications of MDA in academic writing (e.g., Gray, 2015; Gardner & Nesi, 2013). Results from the situational analysis show that composition writing most closely resembles the social sciences and arts and humanities. For instance, assignments in composition, social sciences, and arts and humanities rely on secondary data to a greater extent, whereas assignments in life and physical sciences focus on giving procedural recounts to a greater extent. The MDA of linguistic features reveals that composition aligns with social sciences on several dimensions. For example, in Dimension 1, most disciplines are characterized by informational production, but composition, philosophy, and education exhibit features of involved production. However, this pattern does not hold across all dimensions; in Dimension 3, composition is more similar to physics and engineering in its reliance on situation-dependent reference. These findings highlight both alignments and divergences between composition and disciplinary writing, offering insights that may help instructors better prepare students for the demands of disciplinary writing, offering insights that may help instructors better prepare students for the demands of disciplinary coursework.

#44

Profiling the L2 Turkish lexicon: a learner corpus approach to CEFR-based vocabulary lists

Yigit Savuran, Stefanie Wulff

The rise in the number of Turkish as a second language (L2) learners (Council of Higher Education in Türkiye, 2025) has created an urgent need for empirically-grounded pedagogical resources. However, there is a significant lack of research based on authentic learner language. To address this gap, we introduce the Turkish Learner Corpus (TURLEC), a ~104,000-token corpus (Savuran & Wullf, 2025) of written and spoken texts from university-level L2 Turkish learners from 206 university-level L2 Turkish learners representing 56 nationalities and 43 different L1 backgrounds. This paper presents the first comprehensive, CEFR-based vocabulary profile of L2 Turkish derived from this corpus.

Our methodology employs a multi-criteria approach to assign each of the 3,584 lemmas in TURLEC to a proficiency level (A1-C1). A lemma's level is determined by triangulating key metrics: (1) range (the number of unique learners); (2) first appearance; and (3) frequency. This learner-centric data is then cross-referenced with a lemma's rank in the Turkish National Corpus for external validation. To finalize the assignments and ensure reliability, this data-driven process was double-coded by a colleague, achieving a high level of inter-rater agreement.

Preliminary findings reveal distinct lexical profiles for each CEFR level and demonstrate that a lemma's "peak usage" level (where its range is highest) is a more reliable indicator of mastery than its first appearance. The resulting L2 Turkish Vocabulary Profile offers an unprecedented, data-driven resource with direct implications for curriculum design, materials development, and proficiency assessment, providing a new empirical foundation for the field of L2 Turkish pedagogy.

References

Council of Higher Education in Türkiye (2025, September 10). Student Statistics. Retrieved from: https://istatistik.yok.gov.tr/

Savuran, Y., & Wulff, S. (2025). Developing a Vocabulary Profile for Turkish L2 Based on Learner Corpus. https://doi.org/10.17605/OSF.IO/BNV3P

#48

Corpus-based identification of Mandarin dative construction prototypes using random forests

Shengyu Liao, Stefan Th. Gries, Stefanie Wulff

In cognitive-linguistic studies, the prototype of a category is commonly defined as an abstract representation comprising an aggregate of features with high cue validity for that category (Bernaisch et al., 2014; Divjak & Arppe, 2013; Gries, 2003). Building on this definition, Gries (2003) developed a multifactorial, corpus-based method to identify the most prototypical instances of a category and to quantify the degree of similarity among its members. Applying linear discriminant analysis (LDA) to the English dative alternation, he computed discriminant scores for each instance of the two constructions and arranged them along a continuum between two extremes. The two instances with the two most extreme opposite scores were identified as the most prototypical examples of the ditransitive and prepositional-to dative constructions. This approach has since become standard practice, employing various statistical techniques such as regression modeling (e.g., Deshors, 2014; Divjak & Arppe, 2013) and tree-based approaches (e.g., Bernaisch et al., 2014).

However, a recurring issue in these studies is that prototypes are usually defined as abstract configurations of features with (the) high(est) cue validities for the category in question, i.e. here each construction. Observed corpus data may simply not contain the combinations of features a classifier/model identifies as most strongly predictive of, and thus prototypical for, each of the constructions in question. In Gries (2003), for instance, the most prototypical ditransitive instance was "going round beer festivals gave [me]NP_Rec [the idea of doing it for a living]NP_Theme", which contains a short, pronominal, human Recipient NP with given information and a long, complex Theme NP with new information, but the definiteness of the Theme NP is not compatible with what the classifier would lead one to expect. Thus, restricting one's identification of prototypical uses of constructions to those observed in actual usage can lead to contradictory results.

To address this limitation, we employed hypothetical data. We generated instances for five Mandarin dative constructions with diverse combinations of grammatical features. A random forest classifier (Gries, 2021) was applied to this hypothetical dataset to identify the theoretically 'perfect' prototypes for each construction. We then assessed which of these ideal prototypes are actually attested in real language use by comparing them against a corpus of authentic Mandarin, drawn from the CallFriend (Canavan & Zipperlen, 1996) and ToRCH2014 (Xu, 2017) corpora.

We conclude by discussing the methodological advantages of this approach, along with its cognitive and pedagogical implications.

#49

Stability and usability of word lists based on forms, lemmas, flemmas, and word families in a specialized corpus

Brett Hashimoto, Elizabeth Hanks, Kyra Larsen, Jesse Egbert

In recent years, there has been proliferation in the number of specialized word lists developed for language learning (see, e.g., Tong et al., 2025). Specialized word lists are most beneficial to L2 learners when organized according to the most important words representing the target language domain. Existing lists are based on different operationalizations of a "word" (i.e., based on orthographic form, lemma, flemma, word family); however, previous research has disputed the benefits of these operationalizations (e.g., Webb, 2021). The present study evaluates the stability and usability of specialized word lists developed across various operationalizations of a word. First, we present a framework for evaluating specialized word lists that encompasses a list's validity, representativeness, stability, coverage, and usability. We propose methods for evaluating each of these dimensions, but the present study focuses primarily on the dimensions of stability (i.e., the extent to which rates of occurrence for items in a list remain stable across corpus conditions) and usability (i.e., the extent to which items in a list may usefully applied in a language learning context).

Stability is evaluated here by assessing whether the frequencies of the words in a list are robust when extracted from randomly resampled corpora. This is done by using 1,000 random bootstrapped samples of a corpus (see Hanks et al., 2024). The samples were taken from the 2,700 text, 48-million-word Corpus of English Business Contracts (Hanks et al., 2024) where we tested the stability of word rankings across word type operationalizations. 95% confidence intervals for word rankings across word type operational conditions were produced and compared. This procedure revealed that word families were the most stable, with confidence interval ranges less than <100 within the first 3,000 word families in a list. The same CI ranges were only found within the first 1,500 words for lemmas and flemmas and the first 1,000 words for orthographic forms.

In terms of usability, qualitative analysis of semantic word meanings within each word type operationalization revealed potential challenges of utilizing lists based on word families in the L2 classroom, such as various word meanings that are not necessarily highly related (e.g., "act", "action", "actionable", "unactioned", "actor", "actress") being subsumed under a single word family, and therefore, within a single word count. We also show that the rates of occurrence for many words using the word family approach are the result of multiple word types with only modest frequencies being combined to produce a high total frequency, despite the meanings of these words not being transparently related through their morphology alone.

These results indicate that word families produce more stable results overall, yet they may still not be the most beneficial to use for language instruction. On the other hand, using orthographic forms ignores the relationships that words have with inflectional morphology (e.g., singular vs. plural nouns). We propose that lemmas produce a balance of stability and usability. We discuss implications of these findings for vocabulary researchers, materials developers, and language teachers/learners in creating, evaluating, and selecting word lists.

#50

Register as a predictor of argument structure construction use

Jesse Egbert, Hakyung Sung

Lexicogrammatical variation is a primary focus for two distinct research traditions: (1) the usage-based constructionist approach (UBC; Goldberg, 2019) and the text-linguistic approach to register variation (TxtLx; Biber, 2019). The UBC approach typically has a micro focus on explaining how individual constructions, or "learned pairings of forms and functions" (Goldberg, 2019: 2), are stored, processed, and produced. In contrast, the TxtLx approach usually has a macro focus on describing variation or covariation in the use of (lexico)grammatical structures across texts and registers.

Both traditions seek to better understand lexicogrammatical variation in language use. The UBC approach, for instance, has highlighted the functional properties of constructions in relation to register (Antonopoulou & Nikiforidou, 2011; Goldberg & Suttle, 2010), yet corpus-based UBC practice has given very limited attention to cross-register variation. By contrast, the TxtLx approach foregrounds variation but tends to abstract away from the functional unity of individual form-meaning pairings. We propose that cross-pollination between these research traditions will yield more fruitful research on lexicogrammatical variation (Kerz & Wiechmann, 2015; Casal et al., 2022). To that end, in this study we analyze how argument structure construction (ASC) use is affected by register, in terms of their prevalence and their contingencies, or the statistical associations between ASCs and verbs (Ellis, 2006). Specifically, we have the following two research questions:

1. To what extent do ASCs—and ASC categories—vary across registers?
2. To what extent are the contingencies between verbs and ASCs dependent on register?

We use Mini-CORE, a sample of 200 texts randomly sampled from each of ten registers in the Corpus of Online Registers in English (CORE; Egbert et al., 2015). Each of the 2,000 texts was tagged using the ASC Tagger (Sung & Kyle, 2024). We quantified occurrence rates for eight active voice ASCs in each text. Addressing our first question, we found that some registers, such as song lyrics, use ASCs much more than others, such as research articles. ANOVAs revealed statistically different ASC distributions across registers. For example, whereas more involved, oral registers (e.g. discussion forums, interviews) use relatively few transitive simple constructions, more informational, literate registers (e.g. encyclopedia articles, news articles) rely on them heavily. Qualitative analyses further showed these register patterns are functionally interpretable.

To address our second question, we extracted the list of verb lemmas for each tag. ASC-verb contingencies (measured with MI, T-score, and DP) for the most prevalent verb lemmas were strongly affected by register. For example, in newspaper articles, the verb say is the verb that is most commonly used in the transitive simple construction, accounting for 10% of the verbs used in that construction. In contrast, the verb say is twenty times less likely to be used in the transitive simple construction in newspaper articles, where it accounts for only 0.5% of all verbs.

We show that usage-based research on ASCs will benefit from accounting for register, and that research on register variation will be more comprehensive if it accounts for ASCs and verb contingencies across registers.

#52

From community-oriented documentation to a socio-linguistically diverse corpus

Irina Burukina, Polina Pleshak, Maria Polinsky

Corpus building is a cornerstone of language documentation, yet creating corpora that are of equal value to local communities and to linguists is not straightforward. This paper presents a project documenting Patzún Kaqchikel (Patzún, Chimaltenango, Guatemala). Kaqchikel is a Mayan language spoken by approximately 410,000 bilingual (Kaqchikel-Spanish) speakers, classified as threatened (Eberhard et al. 2022) or vulnerable (Moseley 2010). Our 2019 project transformed pandemic challenges into opportunities for methodological innovation, ultimately producing a corpus that bridges community engagement with rigorous linguistic analysis and thus demonstrating new approaches to adaptive research design.

The project initially aimed at creating a trilingual (Kaqchikel-Spanish-English) book of recipes and oral histories published collaboratively with the Patzún Women's Cooperative Aj Su'm, a Guatemalan NGO, and a US publisher. From the start, topic selection and interviewee recruitment were done with the help of the Cooperative to ensure local relevance. In 2019, we recorded ten hours of spoken Kaqchikel from 17 speakers, representing different occupational and educational backgrounds.

Rather than attempting broad coverage, we deliberately excluded topics that challenged speakers' memory or poorly reflected contemporary speech patterns, such as traditional songs. Instead, we focused on four key genres: everyday recipes, local customs, modern life, and oral histories from Guatemala's 1990s Civil War. The project initially centered on women's experiences, exploring traditional crafts, small businesses, childbirth, motherhood, and education. We later expanded it to include men's perspectives—for example, by incorporating stories from Mayan and Catholic priests—to capture a fuller picture of daily life in Patzún. The principal investigators (PIs) lived with local families, joining in cooking, crafts, household tasks, and community festivals. This immersive approach allowed us to build trust and encouraged natural interactions. Since speakers knew their stories would reach community members rather than outside researchers, they adopted richer, more engaging narrative styles than typically seen in formal elicitation sessions.

Planned for completion by 2021, the project included a 2020 field trip for collecting additional textual and non-textual materials; however, the COVID-19 pandemic disrupted these plans. As book publication was deferred, we prioritized creating an annotated, open-access online corpus of 7.25 hours of Kaqchikel, transcribed and translated by a trained native-speaker linguist into Spanish using ELAN for audio synchronization. The transcriptions were verified by the PIs, who also translated the data to English. A subcollection of app. 3,500 tokens was glossed in FieldWorks. This corpus expands the existing body of Kaqchikel texts (Bennett & Henderson 2022), by incorporating a wider range of topics and more diverse speaker representation across age, gender, and local varieties, supporting broader comparative approaches to spoken Kaqchikel and code-switching.

Despite pandemic-driven shifts from in-person to online formats, the project demonstrates how aligning documentation with community priorities can produce a linguistically robust, socio-culturally representative corpus with diverse narration styles. By adapting to create an accessible online corpus while ensuring local community access, we developed a lasting resource that both celebrates Patzún heritage and advances academic research on Mayan languages.

#54

“I can’t tell bad stories in Spanish:”comparing the effect of topic valence on lexis in oral vs. written narrations from L2 Spanish learners

Andrea Hernandez, Sophia Minillo, Claudia Helena Sanchez-Gutierrez, Paloma Fernandez-Mira

Pivotal to learner corpus research and L2 assessment, task design plays a key role in shaping learners’ language production (Minnillo et al., 2024; Sánchez-Gutiérrez et al., 2024; Tracy-Ventura & Myles, 2015). However, emotional valence (i.e., the positivity/negativity of a prompt’s topic) is often overlooked as a factor in task design. Previous research shows that emotion can shape language learning and processing (Driver, 2022). Emotionally-charged words, especially negative ones, are often easier to remember than neutral ones (Inaba et al., 2005; Kuperman et al., 2014). Previous Spanish learner corpus research has demonstrated a valence effect on the lexical diversity of learners’ writing (Fernández-Mira et al., 2021). These findings suggest that emotional valence might also shape the language that learners produce in oral tasks, but this question has not been widely explored.

This study investigates whether the emotional valence of a prompt topic affects how L2 learners of Spanish use language in narration tasks, and whether this differs between oral and written task modalities. Specifically, we compare learners’ lexis in response to prompts that differ in emotional valence (i.e., positive or negative). We operationalize lexis in terms of lexical diversity, density, sophistication, and appropriateness regarding the psychological property of word valence (Díez-Ortega & Kyle, 2024).

Our dataset includes 160 texts from learners of Spanish as an L2 at a public university in the U.S. The students were enrolled in Spanish courses at the upper beginner or lower intermediate course levels. We collected 40 texts from each of the following prompt conditions: (1) oral positive valence, (2) oral negative valence, (3) written positive valence, (4) written negative valence.

The written texts were part of the Corpus of Spanish-L2 and Heritage (COWS-L2H; Yamada et al., 2020) and were 250-500 words in length. Oral samples were collected through online interviews with Author1 and lasted approximately 10-15 minutes. The interviews were transcribed following CHAT (Codes for the Human Analysis of Transcripts) conventions, and using the Sonix platform to ensure precision and accuracy. We measured lexis through the following indices:

1. Lexical diversity– Mean Textual Lexical Diversity (MTLD), calculated using Frens’ (2017) python script,
2. Lexical density– ratio of content to function words,
3. Lexical sophistication– log frequency of the content words at the lemma level, calculated using TAALES_ES (Díez-Ortega & Kyle, 2024).

We also assessed the appropriateness of students’ vocabulary usage, considering the prompt they were responding to, using the average valence score of the content word lemmas in their text. Valence ratings were taken from Martínez et al. (2024).

Linear mixed-effects models were used to determine the extent to which students’ lexis differed between the two valence (i.e. positive vs. negative) and modality (i.e. written vs. oral) conditions. Preliminary findings suggest that narrations with negative valence exhibit lower lexical diversity and sophistication, not only in writing but also in oral samples. The presentation will describe results and their implications for corpus design and for drawing empirical and pedagogical conclusions from learner corpus analyses.

#55

From learner corpus to classroom:
how can we translate usage-based SLA findings into pedagogical practice?

Ute Römer-Barron

Usage-based approaches to language that take language (in) use and its impact on language acquisition seriously have become increasingly popular and influential in applied linguistics (e.g., Cadierno & Eskildsen, 2015; Ellis & Wulff, 2020; Tyler et al., 2018). We have numerous insights from learner corpus research in the usage-based tradition on how language input affects second language (L2) acquisition. For instance, we know from this type of research that L2 learners are sensitive to frequency distributions in the input they receive (Ellis et al., 2016), that higher input frequencies facilitate stronger entrenchment of language features (Ellis, 2019), and that input enhancements may facilitate the fast learning of new constructions (Goldberg & Casenhiser, 2008). However, we know little about what exactly these insights could mean for the practice of L2 instruction and how pedagogical materials could be adjusted to better reflect usage-based acquisition principles. While many learner corpus studies within the usage-based tradition point out their relevance for language teaching and include general statements on pedagogical implications, they lack direct links to pedagogical practice and tend to remain vague about how L2 teachers could use the insights they share to support their students.

The goal of this paper is to address this gap and discuss how findings from usage-based second language acquisition (SLA) research could be translated into pedagogical practice. The paper will start with a brief overview of the main determinants of language learning from a usage-based perspective and will then summarize findings from recent learner corpus studies on second language development (with a focus on verb constructions and phrase-frames) which allow us to formulate takeaways about how a second language develops and what affects this development. The paper will then present examples of pedagogical interventions that are inspired by these findings and takeaways. In addition to providing specific recommendations for second language education, the paper will propose sample pedagogical materials for the teaching of constructions in instructed SLA settings that are directly inspired by findings from longitudinal and cross-sectional learner corpus studies. It will also briefly comment on the effectiveness of pedagogical uses of corpora, notably data-driven learning (DDL; Johns, 1991), from a usage-based SLA perspective and highlight the need for future learner-corpus-based research to support language teachers and their students.

References

Cadierno, T., & Eskildsen, S. W. (Eds.). (2015). Usage-based Perspectives on Second Language Learning. Berlin: De Gruyter.

Ellis, N. C., Römer, U., & O'Donnell, M. B. (2016). Usage-based Approaches to Language Acquisition and Processing: Cognitive and Corpus Investigations of Construction Grammar. Malden, MA: Wiley.

Ellis, N. C., & Wulff, S. (2020). Usage-based approaches to L2 acquisition. In B. VanPatten, G. Keating, & S. Wulff (Eds.), Theories in Second Language Acquisition: An Introduction (pp. 63-82). London: Routledge.

Goldberg, A. E., & Casenhiser, D. (2008). Construction learning and second language acquisition. P. Robinson & N. C. Ellis (Eds.), Handbook of Cognitive Linguistics and Second Language Acquisition (pp. 197-215). London: Routledge.

Johns, T. F. (1991). Should you be persuaded—Two samples of data-driven learning materials. In T. F. Johns & P. King (Eds.), Classroom Concordancing. ELR Journal, 4, 1-16.

#58

Mapping disciplinary writing development with multi-dimensional analysis

Haiyin Yang, Stefanie Wulff

Despite the importance of writing skills across various domains of life -- such as college preparedness (Smith et al., 2000), gateway to employment and promotion (National Commission on Writing for America's Families & Colleges, 2004), etc. -- American children, youth, and college students' writing skills are widely regarded as insufficient (Conrad, 2017; Rai & Lillis, 2013). One key aspect in which traditional K-12 and college-level writing instruction appears to lag behind is that students are not taught explicitly how best practices and conventions differ between different genres, disciplines, and in academia vs. industry, thus continuing to ignore the vast body of research that has documented how different disciplines and genres employ different linguistic features, rhetoric structures, and assumptions of readers' knowledge to fulfill disciplinary- and genre-specific communication purposes (Hasan & Williams, 1996); (Fang, 2023). Further complicating the matter is the various language background of learners, including native English speakers, heritage speakers of English, and second language learners of English, and next to no research has been devoted to examining whether, and how much, these different student demographics differ in terms of how their writing skill development unfolds over time.

This on-going research aims to map disciplinary writing development by collecting writing data and utilizing multi-dimensional analysis (MDA). Data include student writing across disciplines and by writers of different language backgrounds (native and L2) -- existing college student writing corpora and student research articles collected by us -- and expert writing in the same disciplines and of similar genres -- successful grant proposals contributed by volunteers, and randomly selected articles from top research journals. The analytical tool, MDA, a quantitative corpus-linguistic method that reduces linguistic features from extensive language data into interpretable dimensions that correspond with register and situational variation (Biber, 1988), will then be applied to the data set to reduce linguistic features into interpretable dimensions. MDA returns a distribution of writing along multiple dimensions, and thus the relative difference between student vs. expert, L1 vs. L2, and texts of various disciplines will be measured along interpretable dimensions using statistical tests (e.g., t-tests, clustering) and modeling (e.g., growth curves).

The outcomes of this study will have implications for the articulation of best practices for college writing instruction that is custom-tailored to the needs of students with different language backgrounds and across different academic disciplines.

References

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

Fang, Z. (2023). Demystifying Academic Reading: A Disciplinary Literacy Approach to Reading Across Content Areas (1st ed.). Routledge. https://doi.org/https://doi.org/10.4324/9781003432258

Hasan, R., & Williams, G. (1996). Literacy in society. Longman.

National Commission on Writing for America's Families, S., & Colleges. (2004). Writing: a ticket to work ... or a ticket out : a survey of business leaders. College Board.

Smith, M. C., Mikulecky, L., Kibby, M. W., Dreher, M. J., & Dole, J. A. (2000). What Will Be the Demands of Literacy in the Workplace in the Next Millennium? Reading research quarterly, 35(3), 378-383. https://doi.org/10.1598/RRQ.35.3.3

#66

Additive discourse markers in academic writing: a corpus-based comparative study

Abdulwahab Alshehri

Research has shown the discourse markers serve crucial functions in academic writing across different varieties of English (Jucker et al., 2010). While substantial research has examined discourse markers within Kachur's Inner and Outer Circle varieties (Gabrys & Gilquin, 2017; Wang, 2020), limited attention has been given to how these features are used in Expanding Circle varieties, including Saudi English (Alshurfa et al., 2022). To bridge this gap, this corpus-based study compares the use of three additive discourse markers (moreover, furthermore, and further) employed in academic writing by Saudi and American scholars.

The study addresses 1) the most frequent additive discourse markers used by both groups and 2) how their usage patterns differ. Following Flowerdew's (2004) methodological framework, a corpus of 24 research articles in applied linguistics (10 Saudi-authored, 14 American-authored) published between 2019-2024 was compiled, totaling 205,435 words. Using Stanford Tagger and AntConc, data was analyzed quantitively (frequency distribution, positional patterns) and qualitatively (rhetorical functions). Findings revealed that 'furthermore' occurred with identical frequency in both corpora, while 'moreover' appeared more than twice as frequently in Saudi texts. Conversely, 'further' occurred almost twice as frequently in American texts. Differences were also found in syntactic positioning and versatility. Rhetorical function analysis suggested that Saudi scholars predominantly used these markers for citation and reference purposes, while American scholars employed them more diversely for statistical reporting, explanation and methodological descriptions.

These distinctive patterns emphasize differences in academic writing conventions between the Expanding and Inner circles, suggesting legitimate linguistic variations rather than errors. This study contributes to World Englishes research and offers implications for English for Academic Purposes, particularly regarding how different discourse conventions reflect authentic scholarly voices across cultural contexts.

#68

Leveraging GenAI for proficiency-aligned feedback on L2 learner writing

Shuyuan Tu

The emergence of generative artificial intelligence (GenAI) models, such as GPT-5, offers the potential capabilities that could recognize and respond to complex patterns that reflect different stages of second language (L2) writing (Kohnke et al., 2023; Steiss et al., 2024; Bonner et al., 2023; Dai et al., 2023; Yoon et al., 2023; Guo & Wang, 2024; Lin & Crosthwaite, 2024). While previous studies have explored GenAI's ability to provide feedback to L2 learner writing, no research has systematically investigated whether GenAI can adjust its feedback based on learner proficiency or align it with established developmental trajectories of grammatical complexity. This gap is particularly important in light of research demonstrating that grammatical complexity develops along predictable stages, moving from clausal to phrasal structures (Biber et al., 2011; Biber & Gray, 2016). Furthermore, research comparing GenAI's feedback with human raters is critical in understanding whether GenAI can adequately align with the developmental needs of L2 learners. The present study addresses these gaps by investigating GenAI's ability to provide proficiency-aligned feedback following Biber et al.'s (2011) grammatical developmental framework.

To ensure a manageable scope for manual revision, 450 learner texts written by L1 Mandarin Chinese learners of English were randomly selected from the EF_Cambridge Open Language Database (EFCAMDAT; Geertzen et al., 2013). The texts are evenly distributed across three CEFR levels (A2, B1, and C1) and account for 50,252 words. GPT-4o, accessed via the OpenAI API, was prompted to give direct error corrections (Ellis, 2008) and generate revised versions of the learner texts, adapting Lin and Crosthwaite's (2024) structured prompt. IELTS Task 2 band descriptions were provided as the revision rubric. Six experienced raters were recruited and trained to revise the same learner essays following the same rubric and instructed to focus on direct revisions of learner writing. Their revised versions served as the benchmark for evaluating GPT-generated feedback.

The Lexicogrammatical Tagger (LxGrTagger; Kyle et al., 2025) and a customized script based on the operationalization of the Developmental Complexity Tagger (Gray et al., 2019) were used to preprocess all texts across the three datasets (learner writing, GPT-revised versions, and human-revised versions). Statistical analyses, including descriptive statistics, Kruskal-Wallis tests, and paired-samples t-tests, were conducted to compare the use of grammatical complexity features across levels and between GPT- and human-revised texts. The findings revealed a clear developmental trend in grammatical complexity in GPT's revisions, showing statistically significant increases across levels for many mid and late-stage grammatical features in the development framework detailed by Biber et al. (2011). Specifically, GPT preserves simpler structures in lower-level learner writing while incorporating higher-stage grammatical features in more advanced-level writing. Compared to human raters, overall, GPT shows strong alignment in providing feedback for intermediate and advanced learner writing, but tends to oversimplify the writing at the beginning level. These findings suggest that GPT can serve as a personalized writing tutor for higher-level L2 writers, whereas for beginning learners, it tends to provide simplified feedback that overlooks developmental features. The study highlights GenAI's potential in providing proficiency-sensitive writing feedback.

#69

Investigating EFL student writing in the CAWESA: an additive multidimensional analysis

Wesley Acorinti, Marine Matte, Larissa Goulart

Previous additive multidimensional analyses (MDA) have examined how language produced by university students varies across different levels of study (Gardner & Nesi, 2012) or proficiency (Chen, 2019; Kim & Nam, 2019). The results from Gardner and Nesi (2012) and Kim and Nam (2019) show that writing tends to become more informational, explicit, and abstract with higher levels of study and proficiency, although findings for narrative style and overt persuasion remain inconclusive. In contrast, spoken data from Chen (2019) point to greater involvement, narrativity, situation dependence, overt persuasion, and abstraction as proficiency increases. The present study contributes to this body of research by examining dimensions of variation in Brazilian undergraduate student writing in English across eight levels of study. Using the Corpus of Academic Written English Students' Assignments (CAWESA; Matte, 2004), we adopt a register approach (Biber & Conrad, 2019) to analyze situational and linguistic characteristics. For the situational analysis, we account for the communicative purpose of each assignment based on Goulart et al.'s (2022) framework. For the linguistic analysis, we conducted an additive MDA to describe how each level of study loads on Biber's (1988) dimensions. The results for Dimension 1 (informational production) and Dimension 3 (explicit reference) are consistent with previous studies, showing a downward and an upward trend, respectively. These findings seem to suggest that these patterns are explained by the sharp contrast between spoken and written language as well as by the communicative purposes typical of the assignments at each level. Dimension 5 (abstract style), however, does not increase, diverging from earlier findings, including Biber (1988), which demonstrated abstract style to be prototypical of academic prose. Instead, texts in CAWESA display more non-abstract features at advanced levels, a pattern that can be attributed to task types such as personal statements and reviews that predominate in Level 8. In addition, we observe an increase in narrativity (Dimension 2) and a decrease in overt persuasion (Dimension 4), suggesting a shift toward linguistic features more typically associated with general academic discourse (Biber, 1988).These findings can inform revisions to the syllabus and structure of the program where the corpus was compiled to support linguistic development across levels of study.

#71

Prompt effects on L2 English writing

Henrik Kaatari, Taehyeong Kim, Tove Larsson, Ying Wang, Pia Sundqvist

It is widely acknowledged that students' use of grammatical and lexical complexity features varies depending on the context in which a text is produced (Biber, 1998; Goulart et al., 2020). In particular, research into second-language (L2) English production has shown differences in linguistic complexity depending on the prompt or topic used (Johnson, 2017; Yang & Kim, 2020). For example, personal or opinion-based prompts tend to result in a higher degree of lexical diversity as compared to impersonal prompts (Johnson, 2017; Yang & Kim, 2020). While it has been established that prompts may affect the linguistic output, we know less about how L2 students interact with the prompt itself and, for example, the extent to which learners rely on the prompt and what factors may help predict this – topics of considerable importance for L2 instruction. In this paper, we introduce a framework for investigating the specific effect of prompt words on L2 writing. This framework considers (a) prompt dependency (the proportion of prompt words used in the text); (b) prompt variability (the diversity of prompt and non-prompt words used in the text); and (c) prompt engagement (the extent to which multiword sequences included in the prompt are changed or included verbatim).

To illustrate the framework, we use data from the Swedish Learner English Corpus (SLEC; Kaatari et al., 2024), which comprises texts by Swedish high-school students written on the same prompt: "What are the most important aspects for leading a good life?". A subset of the texts have been proficiency rated (using the CEFR scale, Council of Europe, 2020); the corpus also includes metadata about the time students engage in self-initated, so-called extramural English activities (i.e., conversation, gaming, reading, social media and watching). Research has consistently shown that extramural English exposure influences L2 writing (Sundqvist & Wikström, 2015; Verspoor et al., 2011). In this paper, we demonstrate how our framework can be used to investigate whether learners' engagement in different types of extramural English activities influences their use of prompt and non-prompt words. Early results show that prompt dependency decreases as proficiency increases. The results also show that both prompt diversity and prompt engagement decrease as proficiency increases. In addition to this, the results show a positive relationship between the total time spent on extramural English activities and the diversity of both prompt and non-prompt words, but not with time spent on extramural English and prompt dependency. Engaging in social media use and gaming is associated with lower prompt engagement, which may reflect a certain level of independence and self-efficacy among students who frequently engage in these activities.

#73

GPT-based assisted editing of pre-publication academic writing: an additive multi-dimensional analysis

Rogerio Yamada

Non-English-speaking researchers are expected to publish in English, yet many struggle to do so given the challenges involved in writing quality academic English (Ädel & Erman, 2012; Bardi, 2015; Baumvol, Sarmento, & Da Luz Fontes, 2021; Belcher, 2007; Biber & Barbieri, 2007; Cargill & Burgess, 2008; Flowerdew, 2012; Pang, 2010; Ventura, 2024; Wray, 2019). Given that Large Language Models (LLMs) have been extensively trained on academic English, they may be leveraged as a form of English for Academic Purposes (EAP) literacy broker to edit pieces of writing so that they achieve the level of English required for publication (Flowerdew, 2012; Lillis & Curry, 2006). However, evidence from prior research suggests Artificial Intelligence (AI) may not replicate human academic writing lexicogrammar (Berber Sardinha, 2024). To explore whether AI can edit academic writing in a way that brings it closer to a publication-level standard, we collected a corpus of pre-publication articles and prompted GPT to improve the writing with a view to publication. The resulting corpus was then analysed using Multi-Dimensional Analysis (Biber, 1988, 1995); more specifically, we conducted an additive Multi-Dimensional Analysis (Berber Sardinha, Pinto, Mayer, Zuppardi, & Kauffmann, 2019) based on the dimensions of variation from Biber (1988). We compared the pre-publication articles (original and AI-processed) to articles published in quality journals. Both corpora (pre-publication and published) were sampled from works written prior to the advent of ChatGPT, matching the same time period and the same academic disciplines. The pre-publication corpus included 153 articles, totalling 746,066 tokens, while the published corpus contained 195 articles, totalling 1,789,557 tokens. The results indicated that AI-assisted academic writing diverged from human standards of academic English, as it amplified informational production characteristics (Dimension 1), explicit reference (Dimension 3), and abstraction (Dimension 5), while adjusting narrativity (Dimension 2). The effect of AI interventions varied across disciplines. Excessive informational production characteristics appeared less critical for the Health and Biological Sciences, whereas excessive abstraction was a concern for all disciplines except Applied Social Sciences. In general, AI tended to model all disciplines after the Health and Biological Sciences (i.e. “hard” sciences), in terms of informational production, and against the Applied Social Sciences in terms of abstract style. The effect on narrativity (Dimension 2) and argumentation (Dimension 4) varied according to discipline. In conclusion, this study found evidence of AI failing to reproduce human-authored academic English faithfully.

#76

Lexical and syntactic predictors of human-judged readability: an interpretable machine learning analysis of main effects and genre interactions

Youngmeen Kim

Traditional readability formulas (e.g., Flesch-Kincaid, Lexile) primarily rely on surface-level proxies, such as word and sentence length, raising concerns about construct and theoretical validity (e.g., Crossley et al., 2023), as well as their practical applicability in selecting instructional texts that match learners' levels. In this study, the CommonLit Ease of Readability (CLEAR) corpus (4,724 excerpts; Literary n = 2,420; Informational n = 2,304) was analyzed using an interpretable machine learning (ML) model, Elastic Net, trained on fine-grained part-of-speech (POS) and syntactic dependency (DEP) features to predict human-judged readability scores.

The main-effects model, which represents the general predictive power of POS and DEP features, explained approximately 40% of the variance in readability on a held-out test set. Features associated with easier texts included past-tense verbs, plural nouns, nominal subjects, root verbs, and object predicates. Features associated with more difficult texts included past participles, prepositions, adjectives, appositions, direct or oblique objects, and proper nouns.

Building on the main-effects model, a genre interaction model was trained by introducing a binary genre indicator and all POS/DEP-by-genre interactions to test whether these relationships were uniform across text types. Adding genre (Literary vs. Informational) and interaction terms to this model yielded a modest gain in fit ($R^2 = 0.41$), but more importantly, revealed systematic differences in how linguistic features contribute to readability across genres. Several predictors varied in strength or direction by genre, indicating that readability cues are not genre-neutral. Root verb density improved readability primarily in Literary texts but was negligible in Informational texts. Adjectival modifiers increased difficulty in Literary texts but decreased it in Informational texts, which is consistent with adjectives functioning as clarifying qualifiers in the Informational genre, rather than stylistic elaboration. Personal pronouns and determiners primarily supported readability in Informational texts, while plural nouns contributed to readability in both genres, but more significantly in Literary texts. Proper nouns were generally neutral in Literary texts, yet reduced readability in Informational texts. Visualizations of marginal effects and qualitative analysis, accompanied by exemplary excerpts, illustrate these genre effects in context and demonstrate that readers may face distinct processing demands across genres, even when surface length is held constant.

Taken together, the findings demonstrate that the distributions of lexical and syntactic categories account for roughly 40% of the variance in human-judged readability, without relying on length-based proxies, and that genre influences these relationships in theoretically meaningful ways. For educational practitioners, the models provide interpretable cues that can inform text selection and curriculum design. For researchers, the findings motivate expanding the feature set to include discourse-level and psycholinguistic indices, and exploring complementary, interpretable ML models to capture the remaining variance while preserving explanatory clarity.

#77

Language of distress: machine learning and corpus-linguistic analysis of depression, PTSD, and anxiety in online communities

Youngmeen Kim, Ute Römer-Barron

Mental disorders such as depression, post-traumatic stress disorder (PTSD), and anxiety disorders affect millions of people worldwide, severely disrupting individuals' social and emotional functioning and often leading to serious impairments. Many individuals turn to online mental health communities, which provide anonymity options and offer stigma-free (Naslund et al., 2021) and cost-free alternatives for those who may face difficulties accessing conventional medical care (Marshall et al., 2024). Given the rapid expansion of these online spaces, examining the linguistic and thematic dimensions of communication in such communities is crucial for improving understanding and support for these potentially vulnerable populations.

This study investigates how expressions of mental and emotional states are manifested across three Reddit communities (r/depression, r/ptsd, r/anxiety; 150,000 posts, over 33 million tokens). Using natural language processing (NLP), interpretable machine learning (ML), and corpus-analytic methods, this study addresses three questions: (1) how accurately ML models classify disorder-specific texts, (2) what topics are frequently discussed across communities, and (3) what distinctive linguistic patterns characterize each disorder.

The ML analysis, conducted with a Support Vector Machine (SVM), integrated multiple feature sets, including n-grams, TF-IDF, part-of-speech tags (POS), and dependency parsing labels (DEP), as well as advanced embeddings (Modern-BERT and Sentence-BERT). Classification results showed strong performance. The linguistic feature-based model, trained on the concatenated vectors of n-gram, TF-IDF, POS, and DEP features, achieved an accuracy of 81.86%, Word-level BERT reached 82.73%, and sentence-level BERT achieved 84.43%. Additionally, topic modeling (using BERTopic) revealed distinct, prevalent themes across communities. In the depression subreddit, topics included family conflict, social isolation, substance use, and appearance concerns. PTSD discussions centered on trauma re-experiencing (flashbacks, nightmares), dissociation, memory disturbance, and therapy (e.g., EMDR). Anxiety-related posts emphasized workplace stress, academic pressure, coping strategies (e.g., mindfulness), and somatic symptoms such as breathing difficulties.

Linguistic analysis further illuminated disorder-specific language patterns. Depression-related texts featured negatively charged self-references ("I do not," "I feel like"), superlative adjectives ("worst"), and pronoun-heavy ruminations, often anchored in existential or temporal uncertainty. PTSD discourse relied heavily on past-tense narratives, sensory descriptors, and terms tied to traumatic recall ("flashbacks," "nightmares"), reflecting hypervigilance and intrusive memory. Anxiety discourse was marked by negations, hypothetical or conditional constructions ("I wish I could," "might"), and uncertainty markers, emphasizing negative anticipation, uncontrollable fear, and physical symptoms ("heart," "breath"). These findings suggest that depression discourse centers on inward rumination and hopelessness, PTSD discourse focuses on traumatic re-experiencing, and anxiety discourse is characterized by future-oriented worry and somatic manifestations. By integrating ML with corpus-linguistic analysis, this study shows that online mental health discourse exhibits disorder-specific linguistic patterns. Practically, the results support the development of explainable AI systems for mental health monitoring, tailored interventions, and early detection that can provide timely support. The findings also extend our understanding of how individuals cognitively and emotionally construe their mental conditions in online spaces. This interdisciplinary work bridges linguistics, psychiatry, and computer science, showing how language use in online communities can offer critical insights into the human mind.

#83

What's salient? Genre matters!

Amir Zeldes

Entities in discourse are not uniformly salient – some people, places or other things stand out more than others. Consider the following sentence:

(1) Oakland nurtured Maya Angelou and later Amy Tan, who was partly schooled in Switzerland.

Previous work on salience in discourse suggests that a range of factors indicate Oakland is more central here than Switzerland (linear order: Oakland first; grammatical function: Oakland is the subject; syntactic embedding: Switzerland is in a relative clause), and that Maya Angelou is more salient than Amy Tan (again linear order, but also explicit coordination, a temporal sequence marked by adverbial later). While correlations between salience and linguistic features such as ordering (Gernsbacher 1997), subordination (Miltsakaki 2003), subjecthood (Tomlin & Myachykov 2019), pronominalization (Kaiser 2005) and more have been studied in single genre contexts, the impact of genre variation on these cues is poorly understood. For example, the relative salience of Oakland and Maya Angelou may vary if this sentence comes from a travel guide (e.g. to Oakland), a biography (Maya Angelou's) or an essay about American writers (which would make it more likely that both authors will be equally salient).

This paper reports on an English corpus-based study of the impact of extralinguistically defined genres (Smitterberg & Kytö 2015) on linguistic exponents of salience. Using 24 genres and diverse annotations from the Georgetown University Multilayer corpus (GUM, Zeldes 2017), we construct multi-factorial models to predict graded salience scores and examine how these models incorporate genre information and respond to perturbations in their inputs. The methodology relies on multiple summaries of each document, where being mentioned in more summaries indicates a higher level of discourse salience, thereby operationalizing salience as "summary-worthiness". The results show that while many features correlate with discourse salience, exceptions apply across genres: frequently mentioned, human, pronominal subjects are usually among the most salient entities, but not if they are the hypothetical agent in a how-to-guide; prepositional locative adjuncts are usually non-salient, but not in a travel itinerary; and many many more.

References

Kaiser, Elsi (2005). When salience isn't enough: Pronouns, demonstratives and the quest for an antecedent. In Ritva Laury (ed.): Minimal Reference: The Use of Pronouns in Finnish and Estonian Discourse (Studia Fennica Linguistica 12). Helsinki: Finnish Literature Society, 135–162.

Morton A. Gernsbacher (1997). Two decades of structure building. Discourse Processes 23(3), 265–304.

Miltsakaki, Eleni (2003). The Syntax-Discourse Interface: Effects of The Main-Subordinate Distinction on Attention Structure. PhD Thesis, University of Pennsylvania.

Smitterberg, Erik, & Kytö, Merja (2015). English genres in diachronic corpus linguistics. In Sundkvist, Peter, Shaw, Philip, Erman, Britt & Melchers, Gunnel (eds.): From Clerks to Corpora. essays on the English language yesterday and today. Stockholm: Stockholm University Press.

Tomlin, Russell S. & Myachykov, Andriy (2019). Attention and Salience. In Dąbrowska, Ewa & Divjak, Dagmar (eds.): Cognitive Linguistics - Foundations of Language. Berlin & Boston: De Gruyter Mouton, 36–60.

Zeldes, Amir (2017). The GUM corpus: Creating multilayer resources in the classroom. Language Resources and Evaluation 51(3):581–612.

#84

Generative AI for academic writing:
comparing few-shot and zero-shot in table-to-text generation

Yiwen Zheng, Daniel Dixon

Understanding the capabilities of Generative AI (GenAI) in contexts of higher education and academia are critical for developing informed university policies. Recent research on GenAI in applied linguistics has primarily focused on the use of GPT models via prompt engineering (Barrot, 2023; Yang & Li, 2024) to investigate their capabilities to support instructional design (Arum et al., 2025) and second language learning (Xu & Tan, 2024). However, despite the growing interest in their pedagogical potential, relatively few studies have examined how these models perform on specialized academic writing tasks, particularly those involving quantitative research reporting, such as table-to-text generation. The present study investigates (1) differences in GPT's output across two approaches to prompt engineering: the zero-shot approach vs. the few-shot approach. Unlike the zero-shot approach, the few-shot approach is a technique in which the model is provided with several examples in addition to a system prompt to better guide output generation, and (2) investigate the extent to which the generated output differs compares to human-written texts (i.e., cohesion, academic word coverage, and overall quality as judged by human evaluation).

This quantitative experimental research evaluates GPT-4o's table-to-text generation using three corpora. Two of the corpora consist of texts generated by GPT derived from published statistical tables extracted from prominent applied linguistics journals. One corpus was generated using the zero-shot approach and the other with the few-shot approach, which included six table-to-text examples (three t-tests and three descriptive statistics). The third corpus was compiled by extracting all original prose associated with the target tables, excluding those used for training. Each corpus contains 50 texts, resulting in 150 total texts. Five PhD students (fourth year and above) rated each text on content, organization, language, and hallucination, where mean score ratings were calculated after with inter-rater reliability. Additionally, cohesion (i.e., Latent Semantic Analysis sentence-to-sentence similarity scores and logical connectives) and lexical sophistication (academic word coverage) were measured to gain insight into the differences. Group differences were examined using ANOVAs and post hoc comparisons.

Some intriguing differences were found, especially with the human evaluation which showed clear differences across the three conditions. For both t-test and descriptive datasets, human-written texts consistently received lower ratings on several criteria, including content, language, organization, and (perhaps most surprising) hallucination compared to model-generated texts. Prompt engineering (zero-shot and few-shot) achieved comparable results overall, with only minor differences in scores. Statistical analyses confirmed significant differences between human and machine-generated texts but no significant differences between the two approaches to prompting. Analyses of linguistic features showed that GenAI-generated texts achieved consistently high lexical sophistication, aligning with disciplinary expectations, while human texts exhibited greater rhetorical flexibility in cohesion. Results for LSA similarity suggested that only human texts scored lower than few-shot output in descriptive datasets. Overall, these findings suggest that current prompting strategies are effective for supporting specialized academic writing tasks. At the same time, the distinct rhetorical qualities of human writing highlight the suitability of GenAI as a complement rather than replacement of human authorship in quantitative research reporting.

#85

L2 writing in Ghanaian high-stakes exams: a register-functional analysis

Bernard Cassie, Kwaku Osei-Tutu, Shelley Staples

Grammatical complexity is recognized as a central construct in second language (L2) writing research and has been used extensively to examine both writing quality and developmental progression in English (Lan et al., 2019). Recent research shows that grammatical complexity is shaped by register rather than serving as a fixed measure, with text purpose, audience, and context influencing writers' linguistic choices (Biber et al., 2011, 2022). Building on this perspective, the study applies a register-functional approach (Biber et al., 2022) to examine how Ghanaian tertiary-level L2 English students produce writing that meets institutional expectations while engaging with prescribed source texts in high-stakes examination contexts.

The study addresses two questions: (1) Do students differentiate their use of linguistic features based on the situational context of the assignment? (2) Do students vary their writing across text types in ways that align with the registers modeled in the exam materials? By addressing these questions, the study explores whether student writing demonstrates sensitivity to genre and situational conventions or diverges from the models in systematic ways.

Data are drawn from a learner corpus of students' English examination texts. The dataset includes four student-produced text types: argumentative (238 files; 192,101 words), narrative (110 files; 78,248 words), expository (97 files; 67,060 words), and descriptive (28 files; 18,951 words). These texts are analyzed in relation to source materials (two texts from each text type) used by instructors during the course for direct comparison. All texts were processed using the Biber Tagger for part-of-speech tagging, followed by the Biber Complexity Tagger to extract grammatical complexity features. The analysis focused on clausal features such as finite relative clauses, finite adverbial clauses, verb-controlled that-clause complements, verb-controlled to-clause complements, and phrasal features such as adjectives as noun premodifiers, nouns as noun premodifiers, of-phrase as noun post-modifiers, and prepositional phrases as noun postmodifiers. Frequencies were normed to 1,000 words to account for differences in text length and genre distribution, and variables were analyzed using means and confidence intervals.

Preliminary findings show that for phrasal features such as adjectives as noun premodifiers and noun–noun sequences, usage was higher in argumentative texts and lowest in narratives, with non-overlapping confidence intervals indicating meaningful differences between these two text types. For clausal features, finite relatives were higher in argumentative texts than in other text types, while verb-controlled that-clause complements were highest in narratives and least in descriptive texts; in both cases, the means and confidence intervals indicate meaningful differences between the specified text types.

This study contributes to understanding how L2 writers in an African tertiary context negotiate academic literacy demands, particularly in high-stakes settings. By situating Ghanaian student writing within frameworks of grammatical complexity, register, and genre variation, the study provides empirical evidence from an underexplored context. It also examines the influence of source texts on students' written production and the extent to which their writing reflects the modeled registers. The project contributes to ongoing discussions of academic writing development, genre awareness, and writing pedagogy, emphasizing the role of context in shaping L2 writing practices.

#90

Who stops bleeding?: a diachronic study of amenorrheic discourses in The Lancet

Veronica Ma, Jack A. Hardy, Paige Crowl

Discourses surrounding menstruation differ across cultures. In some contexts, menstruation is seen as sacred (Redland, 2020), but more commonly, menstruation is viewed as shameful and dirty (Malson & Ussher, 1996). In societies where menstruation is often heavily stigmatized, what discourses surround amenorrhea, the absence of menstruation? Previous scholarship has explored cultural discourses surrounding amenorrhea, particularly secondary amenorrhea (the extended absence of menstruation in a person who previously regularly menstruated). These studies use qualitative methods to examine the symbolic representation of amenorrheic bodies in history (Redland, 2020), the links between amenorrhea and anorexia (Malson and Ussher, 1996), and the experiences of amenorrheic athletes (Thorpe, 2016). Scholars have also interviewed people with amenorrhea, highlighting their frustration with medical professionals (e.g., Thorpe, 2016).

This study expands upon previous scholarship through a quantitative, corpus-based approach, analyzing medically authoritative discourses surrounding menstruation. To that end, we compiled a corpus of articles from The Lancet, one of the oldest medical journals. We collected articles published from 1950 to 2024 that contain words such as menstrua*, menopaus*, and other terms related to menstrual disorders (6,378 articles; 20,798,347 words). We used AntConc (Anthony, 2022) to extract the top fifty collocates of amenorrhoea for each decade. This diachronic analysis includes a semantic categorization of collocates to better understand how medical discourses of amenorrhea have shifted over time. In the 1960s, discourses around amenorrhoea focused on the disorder's link to weight loss: amenorrhoea commonly co-occurred with anorexia and, in later decades, began co-occurring with weight. Research on athletic amenorrhoea increased in the 1980s when amenorrhoea began to co-occur with words relating to sport (e.g., athlet*, runners). The 1980s and 2000s also featured more articles on "post-pill amenorrhoea" (the continued absence of menstrual periods after taking birth control). We approach this research using Critical Discourse Analysis (CDA) (Fairclough, 1995) to examine the changing dominant (Western, cisgendered, and male) discourses of the traditional medical research community. As such, we also consider the potentially problematic discourses within these articles. For example, to our knowledge, none of the authors in the corpus are amenorrheic (or at least write about their research from that overt perspective). Also, by predominantly focusing on cis-women subjects, medical research excludes trans experiences. We thus call for future research to collect and study discourses created by amenorrheic individuals themselves, from which we can contrast the less personal medical discourses.

References

Anthony, L. (2022). AntConc (Version 4.1.4) [Computer Software]. Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software

Fairclough, N. (2013). Critical Discourse Analysis: The Critical Study of Language (2nd ed.). Routledge.

Malson, H., & Ussher, J. M. (1996). Bloody Women: A Discourse Analysis of Amenorrhea as a Symptom of Anorexia Nervosa. Feminism & Psychology, 6(4), 505-521. https://doi.org/10.1177/0959353596064003 (Original work published 1996)

Thorpe, H. (2015). My hormones were all messed up. In Endurance Running: A Socio-Cultural Examination (pp. 163–180). essay, Routledge.

Redland, D. (2020). Psychoanalytic Perspectives on Women, Menstruation and Secondary Amenorrhea (1st ed.). Routledge. https://doi.org/10.4324/9781003030478

#97

A corpus-based approach to a CEFR-aligned Korean vocabulary profile

Jieun Kim, Yoonseo Kim

While many European languages have benefited from resources aligned with the Common European Framework of Reference for Languages (CEFR), Korean lacks comparable materials. Existing Korean L2 vocabulary lists (e.g., Korean Language Learning Vocabulary List, TOPIK Vocabulary List) are not aligned with proficiency scales and lack empirical validation. To address this gap, we developed a CEFR-aligned KVP containing 3,193 words ranging from A1 to B2 levels. The KVP draws on the first language corpus (Morpheme-Annotated Corpus; National Institute of Korean Language, 2020), the second language (L2) corpus (Korean Learner Corpus; National Institute of Korean Language, 2023), and existing vocabulary lists. The present study aims to empirically validate this CEFR-aligned KVP using corpus data, guided by the following research questions:

RQ1. To what extent does the KVP align with teachers' and experts' judgments of CEFR levels?
RQ2. To what extent do the CEFR levels in the KVP correspond to Korean learners' vocabulary knowledge thresholds?
RQ3. How accurately can CEFR levels be assigned based on linguistic and/or pedagogical features of words?

From the KVP, 380 words were randomly selected for validation involving 18 language experts, 49 teachers, and 81 learners. After a brief training session, language experts and teachers rated the CEFR level, difficulty, and usefulness of each vocabulary item. Learners completed a test assessing word knowledge by providing the meanings of Korean words. To answer RQ1, an ordinal mixed-effects model was constructed, with CEFR ratings as the outcome variable, rater type as a fixed effect, and words and individual raters as random intercepts. For RQ2, an ordinal logistic regression model was developed, with difficulty based on a dichotomous Rasch analysis of learners' responses as the predictor and the KVP's CEFR levels as the outcome variable. For RQ3, machine learning-based classification models were constructed, ranging from the simplest Model 1, which used the sampled 380 words and all available predictors, to Model 5, which included all words from the KVP. These models employed XGBoost and Decision Tree algorithms.

The ordinal mixed-effects model indicated that rater type (KVP vs. experts and teachers) was not a significant predictor, with threshold estimates revealing clear distinctions among CEFR levels. The ordinal logistic regression showed a significant positive relationship between item difficulty, based on learners' responses, and the KVP's CEFR levels. Finally, compared with the baseline model that included all features (participant-related, learning materials, and corpus-related factors), Models 2 and 3, which included only learning materials and corpus-related factors, demonstrated comparable performance.

Overall, the KVP demonstrated substantial reliability, as evidenced by the alignment between sampled KVP words and expert judgments, as well as learner performance. The machine learning approach further indicated the potential to extend coverage to words at the C1 and C2 levels. The process used to develop the KVP offers valuable insights for creating CEFR-aligned vocabulary profiles for other languages. In conclusion, by validating the KVP through input from various stakeholders and demonstrating its predictive utility, this research contributes both practically and methodologically to the field of Korean as an L2."

#100

Text-internal register shifts in web texts - a cross-linguistic approach

Veronika Laippala, Alireza Razzaghi, Erik Henriksson

In the text-linguistic approach to register variation (Biber 2019), registers are seen as culturally recognized text varieties associated with a communicative situation. Registers are analyzed through pervasive linguistic features that spread throughout the text. Although text-internal variation, especially register shifts within documents, have attracted some attention in register studies as well, the internal structure of texts is much more central in discourse-analytic approaches to genre (e.g., Swales 1990).

In a recent study, however, Henriksson et al. (2025) presented a sub-document register segmentation method, showing that text-internal passages featuring distinct registers can be found and automatically identified in web data. In this presentation, we examine the inner structure of texts as reflected by the predicted sub-document registers. How frequent are texts with different registers assigned to different text segments? What typical register sequences can we find, and how do these relate to the overall register of the text? How do these vary across languages?

For our analysis, we use the web-based HPLT dataset covering altogether 50 terabytes of compressed text data in 198 languages from the unrestricted web (Burchell et al. 2025). We focus on a subset of 3 million documents featuring Finnish, English and Persian. The dataset has document-level register information produced with the multilingual register identification tools developed by Henriksson et al. (2024), following the register scheme of the Corpus of Online Registers of English (Egbert et al. 2015). The scheme is hierarchical, with nine main registers, such as Narrative and Informational Description, and 25 subregisters, such as News Article and Forum Discussion. Documents can also form hybrids with several register labels, and they can be left without any register label if they do not correspond to any of the predefined classes.

The text-internal register labels in the corpus are created with the register segmentation method by Henriksson et al. (2025). Our first findings indicate that 26% of the documents have several registers assigned to their subparts. Altogether in the data, the most frequent combinations are Narrative + Informational Description and Narrative + Opinion, both assigned to documents with document-level Narrative labels. This suggests that the document-internal analysis further deepens the information reflected on the document level. In the presentation, we will apply text-dispersion keyness to examine the linguistic characteristics of these segment combinations and the text-internal register organization they reflect.

References

Biber, D. (2019). Text-linguistic approaches to register variation. Register Studies, 1(1), 42–75.

Burchell, L. et al (2025). An expanded massive multilingual dataset for high-performance language technologies (HPLT). arXiv. https://doi.org/10.48550/arXiv.2503.10267

Egbert, J., Biber, D., & Davies, M. (2015). Developing a bottom-up, user-based method of web register classification. Journal of the Association for Information Science and Technology.

Henriksson, E., Hellström, S., & Laippala, V. (2025). Analyzing register variation in web texts through automatic segmentation. In Proceedings of NLP4DH. ACL.

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

#104

From reputation to survival: student discourses of university life in reviews and reddit

Savannah T. Brown, Jack A. Hardy

Concerns about student mental health and wellbeing have intensified, with surveys consistently highlighting stress, anxiety, and barriers to care in U.S. higher education (e.g., Lee et al., 2021). While institutional data and clinical studies capture aspects of these issues, relatively little is known about how students themselves construct and circulate discourses of university life across different communicative settings. This paper uses corpus linguistics to compare student-authored evaluations of universities in two contrasting but complementary arenas: public-facing university reviews and peer-facing online forums (Reddit).

The analysis draws on the American Corpus of University Reviews (ACUR), a large dataset of student-authored reviews from Niche.com, alongside a newly compiled corpus of posts from university-specific subreddits (e.g., r/Emory, r/UCLA, r/Michigan). Together, these sources allow for comparison of how students evaluate their institutions when addressing different audiences: prospective students and administrators in the review genre, versus peers and fellow insiders in the forum genre. This dual perspective draws on audience design theory (Bell, 1984) to theorize how stylistic shifts emerge across communicative contexts and engages with recent work highlighting Reddit as a site of candid peer discourse.

Methodologically, the study follows a corpus-assisted discourse studies (CADS) approach. Using AntConc, keyword analysis identified salient evaluative patterns, supplemented by qualitative concordance and collocation analysis. To capture discourse-pragmatic variation, we annotated speech acts such as complaint, praise, warning, and suggestion, building on frameworks from consumer review research (e.g., Ludwig et al., 2013; Mauri & Minazzi, 2013; Taecharungroj & Mathayomchan, 2019). Particular attention was given to language related to support services and institutional care, while also extending to evaluations of academics, workload, and campus belonging.

Preliminary findings indicate that student discourse in reviews and Reddit diverges in systematic ways. Reviews tend to frame institutions in reputational terms, aligning with consumerist logics that present higher education as a service to be rated and ranked. In contrast, Reddit discussions foreground survival discourse, centering on peer-to-peer advice, informal resource-sharing, and candid institutional critique. While reviews often compress diverse experiences into generalized assessments, Reddit posts reveal more differentiated accounts of navigating barriers, stigma, and alternatives to official resources. These contrasts extend beyond wellbeing into discourses of academic rigor, advising quality, and campus culture, underscoring how institutional life is framed differently when students write for external audiences versus trusted peers.

By juxtaposing public and peer-facing student discourse, this study contributes in two key ways. First, it demonstrates the methodological value of CADS for comparing genres with distinct communicative purposes and audience orientations. Second, it advances understanding of how students linguistically frame and contest their academic and personal experiences across digital environments. The findings highlight tensions between the reputation-oriented framing of higher education in consumer-facing reviews and the survival-oriented strategies of students in peer-facing forums, offering insights into the discursive negotiation of higher education in the twenty-first century.

#105

Enhancing first-year writing instruction at Hispanic serving institutions with corpus-informed pedagogy: resources, strategies and responsiveness

Anh Dang, Shelley Staples, Randi Reppen

Corpus-informed pedagogy has been widely recognized in applied linguistics as a means of fostering genre awareness, language reflection, and writing development (Boulton & Cobb, 2017; Chen & Flowerdew, 2018; Friginal, 2018). However, much of the research on corpus-informed teaching is focused on the highest research university contexts with extensive resources (Boulton & Cobb, 2017), leaving community colleges and teaching-focused institutions underrepresented. To broaden our understanding of corpus-informed pedagogy across diverse contexts, this study examines its use in community colleges and non-R1 universities designated as Hispanic-Serving Institutions (HSIs). Enrolling large numbers of multilingual and first-generation students while often operating with limited resources, these institutions are critical yet understudied sites. By focusing on four different HSI community colleges and teaching-focused universities in Arizona, this research underscores the importance of context and student population in shaping how corpus-informed pedagogy can be adapted, sustained, and made meaningful.

For this project, we draw on the Corpus & Repository of Writing (Crow), a learner corpus of 18,721 student-authored texts (18,612,586 words) from first-year composition courses at the University of Arizona, Purdue University, and Northern Arizona University, two of which are HSIs. Collected between Fall 2009 and Summer 2022, these texts represent students from over 50 countries and more than 100 majors and include genres such as literacy narratives and research-based arguments. By using this unique corpus of student writing, which enables analysis of language variety and provides student-authored "mentor texts" for instruction (Reppen, 2010), we focused on how corpus-informed pedagogy can foster genre awareness, rhetorical reflection, and informed language choices in ways that support linguistically responsive pedagogy. To support instructors in implementing this pedagogy, we developed a fellowship model that combined professional development workshops, collaborative materials design, and reflective inquiry. The program design was guided by Kolb's experiential learning cycle (1984) and the TPACK framework (Mishra & Koehler, 2006), ensuring integration of content knowledge (corpus content knowledge and first-year writing content knowledge), pedagogical strategies, and technical knowledge with corpus tools. A total of nine instructors from four different HSIs in Arizona participated in our study across two fellowship cohorts. Data included instructor reflections, focus groups discussions, and end-of-program surveys and were analyzed qualitatively using thematic coding, with attention to how instructors navigated opportunities and challenges in adapting corpus-informed pedagogy. Thus, we seek to answer two research questions:

1. What resources and strategies most effectively supported instructors' implementation of corpus-informed teaching at HSIs?
2. How did instructors demonstrate responsiveness to student needs when adapting corpus-based activities?

Preliminary findings show that model lesson plans, peer collaboration, and facilitator support were key in lowering barriers to corpus use. Instructors adapted corpus-informed tasks to validate multilingual practices, connect to ongoing writing projects, and foreground student voices. These results suggest that corpus-informed pedagogy, when scaffolded through accessible training and locally relevant materials, can function as a sustainable, equity-oriented practice in under-resourced institutional contexts. This study contributes to corpus linguistics research on teacher professional development by demonstrating how corpus-informed pedagogy can extend to the diverse contexts of HSIs where many multilingual students learn to write.

#107

Linguistic variation in identity-first vs. person-first language among autism community stakeholders in a reddit corpus

Alyssa Lowry, Earl Brown

Background: For various communities, there exists competing variation between identity-first language (IFL; e.g., autistic person) and person-first language (PFL; e.g., person with autism) preference (Dunn & Andrews, 2015). Existing literature has commented on IFL vs. PFL usage regarding various conditions and disabilities (Jensen et al., 2013; Dunn & Andrews, 2015; Bednarek & Bray 2023; Taboas et al., 2023). However, none have applied a corpus linguistics approach to studying IFL and PFL variation regarding the autism community. Survey data shows that typically, autistic individuals prefer IFL and other stakeholder groups, such as parents and professionals, prefer PFL (Taboas et al., 2023). However, it remains unclear how IFL and PFL vary in frequency and form. For the present study, 11,489 posts (including comments) were scraped from 16 autism-related subreddits to create a 13-million-token Autism Reddit Corpus.

Methods: The corpus was part-of-speech tagged using spaCy's (Honnibal et al., 2020) web-based English transformer model (en_core_web_trf; 2024) and semantically tagged using Python Multilingual Ucrel Semantic Analysis System (PyMUSAS; Moore & Rayson, 2022). Queries to identify virtually all instances of prototypical forms (compare with canonical forms in Bednarek & Bray, 2023) of IFL and PFL (e.g., autistic NOUN and NOUN with autism) were performed using regex in python; search terms also included semantic tags for human nouns. Exploratory concordance analyses were performed in LancsBox X (Brezina & Platt, 2024) to find non-prototypical forms of IFL (e.g., autistic(s) as a noun, audhder(s)). Queries to identify frequency of noncanonical forms of IFL were performed using regex in Python.

Results: Results showed that IFL (1292.05 per million) occurred more frequently than PFL (117.43 per million) across all subreddits. Additionally, PFL occurred significantly more frequently in subreddits relevant to parents of autistic individuals (two subreddits) than in subreddits most relevant to autistic individuals themselves (14 subreddits). Variation between the prototypical form of IFL (autistic NOUN) and less frequent forms (e.g., autistics and audhder(s)) across subreddits will be discussed.

Conclusions: These findings contrast somewhat with previous survey and forced-choice task data regarding stakeholder preferences (Taboas et al., 2023), which suggests that preference and usage do not consistently align for stakeholder groups that report a preference for PFL. However, frequency data seems to validate previous claims that IFL flows more naturally than PFL (2013) and that PFL is an "awkward" and "unconventional style of language" (Taboas et al., 2023, p. 1). Followup analyses with additional subreddits to explore data from more autism stakeholder groups will be discussed, as well as a comparison of IFL and PFL variation between the autism and Asperger communities. The methods in this study can be applied to other communities and platforms to explore IFL and PFL variation as it relates to preference, usage, and identity within specific communities. Furthermore, such an expansion of the current methodology to other communities would allow for comparisons of IFL and PFL variation across communities, speaker groups, and languages, providing cross-linguistic evidence for broader inferences to be made.

#110

Self-mention in discussion sections: a corpus-based comparison of single and co-authored research papers from applied linguistics journals

Juan Rostrán Valle

Substantial research has investigated the use of self-mention in academic writing, with particular attention to writer identity, authorial voice (e.g., Kuo, 1999; Hyland 2001; Hardwood, 2005; Matsuda, 2015; Wu & Zhu 2014), and disciplinary variation (e.g., Hyland, 2001; Hyland, 2002; McGrath, 2016; Tao, 2021; Wang & Hu, 2023). In applied linguistics, studies have primarily focused on self-mention in research article abstracts and some have examined differences between qualitative and quantitative papers (e.g., Ash'ari et al., 2023; Dobakhti & Hassan, 2017; El-Dakhs, 2018). However, distinctions in the use of self-mention in single-authored and co-authored papers warrants further exploration. This corpus-based study investigates the frequency and rhetorical functions of self-mention markers in the discussion section of research articles from three high-impact academic journals in applied linguistics: TESOL Quarterly, Journal of Second Language Writing, and Computer Assisted Language Learning. A total of 150 published research articles, 50 from each journal, were collected and analyzed using a mixed-methods approach. The quantitative analysis, conducted using AntConc version 4.2.4 (Anthony, 2023), examined the overall frequency of self-mentions. The qualitative phase analyzed the distribution and rhetorical functions of these markers in single-authored versus co-authored papers, following an established taxonomy (e.g., Yang & Allison, 2003). Findings show that self-mentions appear across all three journals but in varying proportions: TESOL Quarterly and Journal of Second Language Writing featured twice as many self-mentions as Computer Assisted Language Learning, suggesting that the journal orientation may influence authorial choices. Authors used self-mention markers for a range of rhetorical functions in their discussion texts, with the highest frequencies occurring in moves such as commenting on results and making deductions. These choices appeared to be influenced not only by rhetorical function but also authorial strategy. Authors use self-mentions to assert stance, claim ownership of contributions, and enhance rhetorical clarity and persuasiveness. In doing so, they challenge the conventional norms of academic writing to imprint their authorial identity. Pedagogical implications for writing instruction will be discussed in this paper presentation.

#111

Measuring lexical complexity in L2-Korean writing through a morpheme-aware approach

Hakyung Sung, Gyu-Ho Shin Shin

In learner corpus research, lexical complexity has long been recognized as an indicator of second language (L2) writing proficiency. While it has been extensively investigated in L2 English, the construct is underexplored in typologically diverse languages such as Korean, with its agglutinative morphology. To address this gap, we draw on a corpus of L2-Korean writings to develop morpheme-aware lexical complexity indices and evaluate their relationship to writing proficiency, using both descriptive analyses and correlation-based tests of predictive power.

To design the experiment, we considered two key factors. First, it was important to establish a clear operational definition of what constitutes a lexical item in order to examine lexical production (Jarvis & Hashimoto, 2021). Whereas in English lexical items are typically defined as words segmented by whitespace and punctuation, we operationalize them at the derivational morpheme level, that is, by segmenting words into content and functional morphemes in analyzing Korean texts. This approach is theoretically motivated by Korean's reliance on case particles and layered verbal morphology to encode grammatical and pragmatic distinctions (e.g., topic vs. nominative marking; tense/aspect/mood and honorific markers) and has been empirically validated as offering finer granularity for reliably capturing lexical phenomena in Korean.

Second, we designed lexical complexity indices based on Bulté and Housen's (2012) tripartite framework—diversity, sophistication, and compositionality, while incorporating insights from previous studies. Lexical diversity was conceptualized in terms of variety, evenness, and dispersion based on previous studies (Kyle et al., 2021; Sung et al., 2024). Lexical sophistication was operationalized as the proportion of low-frequency words relative to a reference corpus, following common practice in L2 English research (Kyle & Crossley, 2016). Finally, lexical compositionality, which concerns the internal structure of words and their morphemic combinations, has often been referred to as morphological complexity. Although this construct has received limited attention in English due to its relatively simple inflectional system (Brezina & Pallotti, 2019), Korean's agglutinative morphology offers a unique opportunity to analyze morpheme compositionality by examining whether words are formed from frequent or infrequent morpheme combinations—an area that has not yet been actively studied.

In the experiment, we analyzed 600 L2-Korean essays using 17 lexical complexity indices. Frequency-based measures were computed with reference frequency norms derived from a 347-million-word L1-Korean corpus. Correlation analyses revealed small-to-medium associations between these indices and proficiency scores, and a regression model with seven predictors accounted for about 36% of the variance in proficiency. Roughly 75% of the model's predictive power came from indices of lexical compositionality, particularly those capturing case-marked nouns and complex predicate endings. This suggests that the morphological markers such as case particles and predicate endings play a critical role for L2-Korean writing. The remaining variance was primarily accounted for by a lexical diversity index, suggesting that breadth of produced morphemes also contributes significantly. Overall, the results demonstrate the value of morpheme-sensitive measurement practices in typologically diverse languages and point to pedagogical implications: expanding learners' morpheme repertoires and strengthening control of high-frequency morphosyntactic forms and combinations to enhance writing competence in L2 Korean.

#112

A lexicogrammatical analysis of a phrasal complexity feature

Elizabeth Meyr, Brett Hashimoto

The predictive power of phrasal and lexical complexity for writing proficiency has been broadly explored in second language academic writing albeit separately (e.g., Biber et al., 2011; Kim, 2014). Studies that do consider both complexity strands often do not include lower-level learners or a variety of first languages (Staples et al., 2023). Attributive adjectives are among the phrasal complexity features that commonly occur in the written academic register (Biber, Gray & Poonpon, 2011). Also, currently there are no studies to our knowledge that investigate one specific feature of phrasal complexity, but this pursuit is valuable to the study of complexity because many studies show varying patterns of usage for specifically the development of attributive adjectives (Larsson et al., 2023; Seo & Oh, 2024; Staples et al., 2023). The purpose of the present study is to evaluate the extent that incorporating lexical complexity into phrasal complexity measures increases the predictive power of proficiency in a learner corpus of ~7,000 written essays from an Intensive English Program (IEP). A linear mixed effects analysis examined the relationship between writing proficiency measured using MFRM-based scores and the normalized frequency of attributive adjectives in a basic model, as well as in a model that incorporates both phrasal complexity and lexical complexity measures. Lexical diversity will be operationalized through the number of different attributive adjectives (NDW) used in each text. Lexical sophistication is realized as academic COCA frequencies of attributive adjectives within each text. The models were compared respectively to explore whether the explanatory power would increase by modifying the lexical complexity measures. The results indicated that incorporating lexical complexity measures within the grammatical complexity model increases the predictive power of proficiency for this phrasal complexity feature, especially NDW ($\Delta R2 = .1$). The models that include diversity measures explain the variance of proficiency scores better than rate of occurrence or the sophistication measures used in this study. This shows that lexical diversity is highly indicative of proficiency in an ESL context, so there are implications for emphasizing less lexical repetition in the classroom. Interestingly, lexical sophistication measures are not as predictive as diversity, which could be influenced by the task and operationalization. Foremost, normed rate of occurrence does not strongly account for the variation within the instances, so it supports previous studies that have shown a weak relationship with proficiency for this feature. This study has implications for revisiting current lexical complexity measures, such as MTLD and MATTR, which have difficulty accommodating for shorter texts, so methods should be evaluated for lexical complexity further on the phrase and clause level. Considering multiple strands of complexity is crucial to understanding the usage of grammatical complexity features, and more features should be evaluated accordingly in the future. This study can be applied to the patterns of students' written complexity through two strands and ultimately contribute to the development of instructional materials that incorporate grammatical complexity.

#113

Developmental trajectories of noun phrase complexity in L2 English academic writing: frequency and lexical realizations

Taehyeong Kim

Based on the register-functional approach to grammatical complexity (Biber et al., 2022), noun phrase complexity has been identified as a marker of advanced second language (L2) English writing (e.g., Wang & Jiang, 2024). While prior studies have consistently supported a developmental shift from clausal to phrasal elaboration in L2 writing (e.g., Larsson et al., 2023), most have relied primarily on frequency of use as an index of development, often overlooking the question of how these grammatical features are lexically realized across stages, especially for constructs such as productivity (type count of lexical realizations) and diversity (type-token ratio of lexical realizations). However, insights from previous studies of grammatical complexity suggest that development also occurs via variability in the lexical choices associated with a grammatical feature, as evidenced by the repeated use of lexical items for a given grammatical feature by L2 writers in comparison to L1 writers (Gray et al., 2023; Lan & Sun, 2019; Staples & Reppen, 2016) and lower-level L2 students' reliance on repetition in prompt-induced adjective-noun sequences compared to advanced L2 students (Seo & Oh, 2024), thus resulting in more limited variability in their lexical realizations.

This remaining question ties closely to a growing view in applied linguistics that lexis and grammar are deeply interconnected (e.g., Römer, 2024). Thus, this study adopts a novel approach incorporating both frequency and the lexical realization of two noun phrase features found to be particularly frequent: adjective-noun (i.e., attributive adjectives) and noun-noun (i.e., premodifying nouns) sequences. To track development, this study examines a corpus of L2 academic writing sampled from academic writing corpora ranging from undergraduate, graduate, and professional levels, given that many previous studies examining the development of noun phrase complexity in L2 English academic writing have focused exclusively on undergraduate and/or graduate contexts. While such work has provided valuable insights into early and intermediate stages of academic writing development, this scope leaves unanswered the important question of whether growth in noun phrase complexity plateaus or continues beyond the undergraduate and graduate years. This study analyzes normalized frequencies of these features, applies moving-window measures to capture productivity and diversity, and uses mean values and non-overlapping 95% confidence intervals to identify meaningful developmental changes (see Staples et al., 2022).

Findings show that frequency and productivity remain stable during undergraduate years but increase at graduate and professional levels. In contrast, diversity patterns differ; adjective-noun sequence diversity increases only at the professional level, while noun-noun sequence diversity, especially in head nouns, shows a gradual progression beginning in late undergraduate levels. These results suggest that L2 academic writing development is not fully captured by frequency of use alone. Lexical realization adds a critical dimension, showing how L2 writers develop lexical control and discourse precision through the interaction of frequency, productivity, and diversity in noun phrase construction.

#115

Expanding the construct of grammatical complexity: the case for grammatical diversity in writing

Christian Holmberg Sjöling, Taehyeong Kim

Much research using an automated and quantitative approach to investigate grammatical complexity features in writing do so with a frequency-based approach (e.g., Biber, 1988; Biber et al., 2011; Biber et al., 2016; Kyle, 2016; Lu, 2010). In this paper, we build on the frequency-based approach to tap into a complementary dimension of grammatical complexity that we term grammatical diversity. Writers who include more diverse grammatical complexity features have a broader productive knowledge of grammatical complexity (i.e., broader grammatical diversity) while writers that produce a smaller set of grammatical complexity features or repeat a small set of features have a narrower knowledge of grammatical complexity (i.e., narrower grammatical diversity) (cf. Jarvis, 2013).

We have written a Python script that builds on tagging initially carried out with the Lexicogrammatical Tagger (Kyle et al., 2025). The script counts the number of unique grammatical complexity features for a moving five-sentence window of a text (i.e., sentence 1–5, 2–6, 3–7 and so on) (cf. Zenker & Kyle, 2021). This approach captures local variability and patterns of grammatical complexity features used in a text and 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 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 (i.e., non-overlapping CIs and 10%+ increases) 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.

Jarvis, S. (2013). Capturing the diversity in lexical diversity. Language learning, 63, 87–106.

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

Kyle, K., Biber, D., Sung, H., Reppen., R., & Egbert, J. (2025). Lexicogrammatical Tagger. [computer software]. https://github.com/kristopherkyle/LxGrTgr. Accessed September 15, 2025.

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

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.

Zenker, F., & Kyle, K. (2021). Investigating minimum text lengths for lexical diversity indices. Assessing Writing, 47, 100505.

#116

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

Stefan Th. Gries, Nina S. Funke

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 generalized linear (mixed-effects) models. Given the emphasis on corpus-based and quantitative analysis, the retrieval of the alternants in the corpus data of course 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.,

- Grafmiller's (2014) study of the genitive alternation removes "all tokens involving pronominal possessors [...] 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.

#119

Lexical and lexico-grammatical predictors of heritage Spanish bilinguals' productive proficiency: a replication of Kyle and Eguchi (2023)

Nate Cook, Gyu-Ho Shin

This study investigates the extent to which automatically extracted lexical and lexico-grammatical indices predict spoken heritage-language (HL) proficiency in Spanish heritage speaker bilinguals (HSBs). Building on Kyle & Eguchi (2023), who reported that a combination of word-level and multiword indices explained a majority of variance in Oral Proficiency Interview (OPI) scores for L2 learners, we replicate and extend their approach to a corpus of Spanish HSB oral interviews. To this end, we pose two research questions: (1) which individual lexical and contiguous-bigram indices relate meaningfully to OPI scores?; (2) to what extent do combinations of these indices account for variance in OPI outcomes?

Data were sourced from the Chicago-Spanish Corpus (CHISPA, Potowski & Torres, 2023; n = 123; ≈472K tokens), which comprises semi-structured OPI interviews with Spanish HSBs representing Mexican, Puerto Rican, and mixed varieties and multiple generational cohorts. We computed a set of word-level indices (e.g., MATTR lexical diversity; log-transformed reference-corpus word frequencies; mean concreteness from Spanish norms) and contiguous bigram strength-of-association measures (e.g., T, MI, MI2, Delta-P) using lemmatized and POS-tagged transcripts processed via spaCy–Stanza pipelines. Reference frequencies were derived from the Corpus Sociolingüístico de la Ciudad de México (Martín Butragueño & Lastra, 2015) and the Cortés-Torres (2005) Puerto Rican corpus to align with informant varieties. Statistical procedures comprised: (i) outlier removal and log-transformation of skewed indices prior to correlation analyses; and (ii) multicollinearity screening and best-subset linear regression with Bayesian Information Criterion model selection and relative-importance estimation (lmg).

Results showed that several indices exhibited significant associations with OPI scores, but the pattern diverged from prior L2 findings. At the word level, (1) adjective frequency correlated negatively with OPI ($p=0.013$), indicating that higher-proficiency HSBs favor lower-frequency (more sophisticated) adjectives, and (2) overall concreteness correlated negatively with OPI ($p=0.001$), consistent with greater abstractness among more proficient speakers. At the multiword level, directional predictability (deltaP_max) correlated positively with OPI ($p=0.022$), whereas the T measure correlated negatively with OPI ($p<0.001$), suggesting that advanced HSBs produce bigrams that are more directionally predictive and contain less frequent co-occurring words. In model selection, two predictors—T and log adjective frequency—comprised the optimal regression, explaining approximately 27% of variance in OPI scores (adjustedR^2≈0.256), with T as the dominant contributor.

These findings partially replicate Kyle & Eguchi (2023) by demonstrating that lexical and lexico-grammatical features relate to spoken proficiency; however, our results also suggest systematic differences for heritage speakers. While concreteness was the dominant predictor in Kyle & Eguchi's model that explained more L2 proficiency variance, HSB proficiency appears more strongly associated with frequency-based dimensions of adjectives and bigram co-occurrence patterns, possibly reflecting varied exposure, sociolinguistic background, and register-specific usage in naturalistic HL contexts. This study's findings thus underscore both the promise and the limitations of L2-informed corpus-based indices for HL variation and assessment: combined lexical and collocational metrics can contribute meaningful signal to proficiency evaluation, but their diagnostic utility may depend on population characteristics and reference norms.

#122

Beyond the news: genre-diverse annotations for bridging anaphora in English

Lauren Levine, Amir Zeldes

Bridging is an anaphoric phenomenon where the referent of an entity in a discourse is dependent on a previous, non-identical entity for interpretation. For instance, in the following sentences, "There is a house. The door is red," the door is specifically understood to be the door of the aforementioned house, and, as such, is dependent upon the house for interpretation. In this example, “the door” is referred to as the bridging anaphor, and “a house” is referred to as the associative antecedent. These constitute a bridging pair, and the “bridging” occurs when the reader constructs an inference back to the associative antecedent in order to interpret the referent of the bridging anaphor (Clark, 1975). Bridging has been studied from a variety of theoretical perspectives (e.g., Hawkins 1978; Asher and Lascarides 1998; Baumann and Riester 2012), and linguistic resources have been constructed with various schemas and definitions of bridging (e.g., ISNotes, BASHI, and ARRAU RST). While several English resources have been annotated for bridging anaphora, most are small, provide limited coverage of the phenomenon (e.g., only definite NPs), and/or provide limited genre coverage (primarily WSJ news data) (Roesiger et al., 2018).

To fill this gap, we introduce GUMBridge, a new resource for bridging, which contains 2.1k bridging instances over 126k tokens of English, covering 16 diverse genres (including academic writing, courtroom transcripts, and conversations). GUMbridge is constructed on top of a portion of GUM v11, an existing multi-genre corpus of English (Zeldes, 2017). GUMBridge provides broad coverage for the phenomenon of bridging, using information status based criteria to identify instances of bridging, which does not place structural limitations on the manner in which bridging can manifest in a discourse. Additionally, GUMBridge provides granular annotations for the categorization of sub-varieties of bridging, introducing a new schema with 10 subtypes divided into 3 main categories: comparison, entity, and set (with an additional other category). We aim to expand the size of GUMBridge to 228k tokens (the full size of GUM v11). This data will support further linguistic analysis, helping to answer ongoing questions of how speakers interpret and resolve bridging across different domains. Additionally, GUMBridge provides more diverse training and evaluation data for the NLP task of bridging resolution, where bridging anaphora and their associative antecedents are automatically identified.

References

Nicholas Asher and Alex Lascarides. 1998. Bridging. Journal of Semantics, 15(1):83–113.

Stefan Baumann and Arndt Riester. 2012. Referential and lexical givenness: Semantic, prosodic and cognitive aspects. Prosody and meaning, 25:119–162.

Herbert H. Clark. 1975. Bridging. In Theoretical Issues in Natural Language Processing.

John A. Hawkins. 1978. Definiteness and indefiniteness: A study in reference and grammaticality prediction. Journal of Linguistics, 27:405–442.

Ina Roesiger, Arndt Riester, and Jonas Kuhn. 2018. Bridging resolution: Task definition, corpus resources and rule-based experiments. In Proceedings of the 27th International Conference on Computational Linguistics, pages 3516–3528.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. Language Resources and Evaluation, 51(3):581–612.

#125

Communicative co-occurrence in conversation: intra-text patterns of communicative purpose and topic across conversational discourse units in BNC Spoken

Daniel Keller, Marianna Gracheva

Communicative and linguistic variation within registers is now well-documented in register research (Biber & Egbert, 2023): It has been consistently shown that registers are not situationally or linguistically uniform, and linguistic variation within registers systematically corresponds to communicative variation among their texts (e.g., Egbert & Gracheva, 2023; Wood, 2023; Goulart, 2024). Most studies of intra-register variation to date have focused on communicative purpose, "the motive force" (Brown & Fraser, 1979:39) or "the 'why' of communication" (Biber & Conrad, 2019:45). Indeed, purpose has produced interpretable patterns of variation among texts of several domains: online registers (Biber & Egbert, 2018), conversation (Biber et al., 2021), student essays (Goulart et al., 2022), university textbooks (Egbert & Gracheva, 2023), legal statutes (Wood, 2023), and even texts without a recognized register category (Egbert et al., 2024).

While these studies, concerned with purpose variation from the start, coded texts for purpose to examine corresponding linguistic patterns, other studies did not begin with a pre-determined situational parameter and aimed to identify bottom-up the basis for the observed linguistic variation among texts (Gracheva, 2023; Author, forthcoming). In these studies, topic emerged as the communicative factor that explained linguistic variation in several registers. [GM1] Taken together, these findings point to a likely unity of topic and purpose in explaining linguistic variation. Because the two factors have often been viewed as inseparable from one another (Brown & Fraser, 1979) and existing frameworks tend to include the specification of subject matter into their purpose label (e.g., "prescribe a mandatory duty or responsibility to a human agent or government entity," Wood, 2023), the respective contributions of these factors into explaining linguistic variation and their relationship remain unclear.

This study aims to examine patterns of co-occurrence between particular topics and purposes, examining their interrelationship and laying the groundwork for subsequent linguistic analyses. We chose a corpus annotated for communicative purpose at the level of the discourse unit—the conversations of the BNC Spoken (2014)—and annotated for topic at the level of the token using the UCREL Semantic Analysis System tagger (Rayson et al., 2004), which assigns each word in each discourse unit to one of 21 discourse fields (e.g., the body and individual, arts and crafts, emotions, food, money and commerce), thus allowing characterization of each discourse unit in terms of its topic(s).

In this presentation, we describe the distributions of major categories of semantic tags across discourse units characterized by dominant communicative purposes like expressing feelings, joking around, or giving advice. We also discuss the results of planned follow-up analyses including Multiple Correspondence Analysis to identify underlying co-occurrence patterns between communicative purposes and topics.

#127

Audiences and discourse types in fiction: a register analysis of children's and adult literature

Marianna Gracheva, Michaela Mahlberg

Although fiction has featured in comparative studies of registers (e.g., Biber 1988), it has largely remained on the sidelines of register studies due to its proverbial situational ambiguity: the situation of fiction is the situation of the fictional world created by the author rather than the external situation of text production (Biber & Conrad, 2019). Therefore, traditional situational frameworks, accounting for factors such as communicative purpose, relationship between participants, and shared knowledge, are not well suited to capture the 'situation of fiction'. Only a few studies have set out to focus on register features of fiction in particular (e.g., Biber & Finegan 1994; Egbert 2012). Egbert and Mahlberg (2020) work with a corpus that divides 19th century fictional novels into the discourse types of narration and fictional speech by separating the language contained in quotation marks from the language outside quotations. They show that narration and fictional speech pattern differently with respect to the language of "Involved vs. Informational Production" (Biber, 1988), "Thought Presentations vs. Description" (Egbert, 2012), and "Dialogue vs. Narration" (Egbert, 2012).

In this study, we build on previous work that separates speech and narration but add a further angle. We examine the discourse types not only with relation to one another but include the text external situational characteristic of audience type. From a literary perspective, the intended readership of a text is worthy of attention (e.g., Rabinowitz, 1977) but from a linguistic point of view this aspect of literary texts has not received much interest. We will specifically look at 19th century novels written for adult and child audiences and work with two of the CLiC corpora (Corpus Linguistics in Context; Mahlberg et al., 2020): a corpus of 19th century fiction (19C, N = 29) and a corpus of children's literature (ChiLit, N = 71). Similar to Egbert & Mahlberg (2020), our method is additive multidimensional analysis (MDA), which plots the corpora under analysis on five fundamental dimensions of variation in speech and writing identified in an earlier study (Biber, 1988). As we apply the MDA, we first compare children's and adult's fiction to each other on the five dimensions—Involved vs. Informational Production, Narrative vs. Non-Narrative Concerns, Explicit vs. Situation-Dependent Reference, Overt Expression of Persuasion, and Abstract vs. Non-Abstract Information; we then examine narration and fictional speech separately: a) discourse types within audience types (i.e., dialogue and narration in children's vs. dialogue and narration in adult's literature); b) audience types within discourse types (i.e., adult and children's literature in dialogue vs. adult and children's literature in narration), to see whether and what the discourse type split adds to the audience comparison. The results show that children's literature is noticeably more involved and elaborated, more situation dependent and concrete, less persuasive, and less abstract. Discourse type differences are large in both children's and adult fiction, but an examination of audience types within each discourse type showed that audience differences are of different magnitude across discourse types. We situate these linguistic findings in the context of literary interpretation.

#132

The relationship between L2 Spanish written narrative retellings and features of lexicogrammatical use

Hana Dussan, Kristopher Kyle, Mery Díez-Ortega, Carla Consolini

Writing proficiency is a particularly important aspect of academic success that is difficult for both first and second language (L2) users to obtain (e.g., Kellog & Bascom, 2007). An important aspect of proficient writing is the appropriate use of lexicogrammatical features. Indices that measure features of lexicogrammatical use (e.g., lexical diversity, lexical sophistication) are consistent indicators of L2 proficiency scores (e.g., Bestgen & Granger, 2014; Kyle et al., 2018, 2021). Accordingly, the development of appropriate use of these features are important in the development of L2 learners' writing competence. Research suggests that more proficient L2 writers tend to use a wider variety of lexical items, more sophisticated lexical items, and more strongly associated word combinations (e.g., Kyle et al., 2018; Kyle & Eguchi, 2021; Sung, 2024; Zenker et al., 2021).

One limitation of this body of research is that it has primarily been concentrated on L2 English learners. However, a recent study conducted by Consolini and Kyle (2024) sought to fill this gap by examining the relationship between features of lexicogrammatical use in L2 Spanish descriptive essays and general L2 Spanish proficiency levels (based on a reading and grammar assessment tool). While the results indicated some links between lexicogrammatical use and proficiency levels, an important limitation was a lack of writing quality scores, which provide a more direct measure of written L2 Spanish proficiency.

This study addresses this limitation via a principled replication of Consolini & Kyle (2024). First, an analytic rubric designed to capture features of lexicogrammatical use was developed and piloted, and raters were trained to use the rubric. Second, raters then rated 325 of narrative retellings from the CEDEL2 corpus (Lozano, 2022). Each essay was rated by at least two raters, and inter-rater reliability was acceptable (ICC = 0.69). Third, indices related to lexical diversity (moving average type-token ratio), lexical sophistication (log-transformed frequency of content words), and the strength of association between contiguous bigrams (adjacent words) and dependency bigrams (e.g., verb-object) were calculated. All corpus-based indices utilized a large web-based corpus (ESCOW; Schafer et al., 2015) as the reference corpus.

The results of a multiple regression analysis using lexicogrammatical indices to predict writing proficiency scores are in line with previous research in L2 English (e.g., Kyle & Crossley, 2017; Bestgen, 2017) and show that lexical diversity scores along with bigram SOA and bigram object dependencies scores explained 29% of the variance in human judgment scores of the essays. In the presentation, data analysis, results, and pedagogical implications will be discussed in detail.

#134

A corpus approach to constructional variation and category change in Portuguese

Chad Howe

The current paper considers the question of change in grammatical category through the lens of synchronic variation, focusing specifically on a class of structures in Portuguese–often referred to as Clausal Existential Constructions (following Rigau, 2001)–in which the verbs haver, fazer, and ter occur with a temporal complement, as in há/faz/tem muito tempo 'a long time ago'. This class of structures is attested across Romance with a wide array of properties (see Howe, 2011, Móia and Alves, 2004, and Rigau, 2001,) that, despite a number of structural and semantic parallels, suggest a neutralizing of the properties of the verbs involved. In this paper, we explain the variable use of haver, fazer, and ter using data from the Brazilian portion of the Corpus do Portugês (CdP, Davies and Ferreira, 2016, Web/Dialects) and find that há (< haver)+time has a wider distribution across contexts, suggesting a greater degree of grammaticalization. Moreover, we provide corpus evidence that, counter to previous claims about the continued status of Spanish hacer+time (see Herce, 2017), the há+time structure, in particular, retains few (if any) vestiges of its erstwhile verbal source.

A sample of tokens was extracted from the CdP using regular expressions in Corpus Query Language (CQL). These tokens were targeted with the verbs haver, fazer, and ter followed by a temporal complement (e.g., muito tempo ‘a long time’, dois anos ‘two years’). The data were coded for a range of structural and semantic features, including (i) position vis-à-vis the modified verb (i.e. pre vs. post-verbal), (ii) presence of a complementizer (in pre-verbal position), (iii) tense (e.g., há vs. havia) and (iv) negation. The tokens were further coded has expressing either a durative reading or a punctual reading. A total 4,372 tokens were analyzed for the current analysis.

As predicted, the most frequently used form, by far (at a 10:1 ratio), was the ha+time collocate. Across the three structures, the results suggest that, while both fazer and ter+time are attested with most of the same structural and semantic properties as há, these collocates are preferred in exactly those contexts that suggest a verbal source, such as in the presence of an overt complementizer in pre-verbal position. Morever, we find há+time in durative contexts modifying a present tense verb, a context that we claim represents a critical development in the attrition of verbal features in this structure. The corpus analysis suggests (i) that the construction with há is more grammaticalized than either fazer or ter and (ii) that structural and semantic changes are not necessarily transmitted uniformly across an entire constructional paradigm. More generally, we maintain, following Michaelis (2010, p. 146), that these structures participate in multiple inheritance hierarchies, which holds that constructions belonging to the same type (i.e. verb, noun, etc.) share properties and that any given construction may belong to more than one type. We build on this view as a basis for a broader proposal regarding category change with these structures.

#135

Linguistic variation and team performance: a corpus study of Brazilian football fans

Alexcia Boothe, Jack Hardy

This study examines how stance and group alignment are linguistically encoded in Brazilian fan discourse on Reddit's r/futebol. Using a 10,000-post corpus (approximately 1.5 million tokens) spanning 2020–2024, we situate the analysis in sociolinguistics and Systemic Functional Linguistics (SFL), focusing on how interpersonal meanings are mobilized in online environments where collective identity and group boundaries are continually negotiated.

We analyzed interpersonal metafunctions by measuring resources for alignment and evaluation across subcorpora representing supporters of high- and low-performing clubs in the 2023 Brazilian Série A Championship. Personal/pronominal reference (including pro-drop cases) was identified by extending NLP tagging with Python/NLTK scripts that infer implicit subjects from verb morphology, improving recall of first-person categories. Team naming practices were retrieved via concordance and keyword analysis in Sketch Engine, normalized to occurrences per 10,000 tokens, and compared across subcorpora. References to agents (players, coaches, referees) were captured through named-entity recognition and collocation analysis to track attribution of agency and blame. Affect and attitude lexis were extracted using Appraisal Theory categories through lexicon-guided extraction and targeted manual checks to ensure polarity/stance accuracy.

Findings show systematic contrasts in interpersonal design dependent on team performance. Supporters of relegation-threatened teams use first-person plural forms more frequently (χ^2 $p < .001$) and refer to fouls/misconduct at higher rates (χ^2 $p < .001$), indexing solidarity under adversity and grievance about officiating. In contrast, supporters of high-performing teams produce more team-name tokens (χ^2, $p < .001$) and use more expressive/evaluative language (χ^2, $p < .001$). They also exhibit relatively higher rates of first-person singular (χ^2 $p < .001$) and third-person reference to players/coaches (χ^2 $p < .001$), consistent with celebratory narration, performance assessment, and individualized commentary. Notably, while top-team fans use third-person pronouns more overall, fans of lower-ranked teams make more direct named references to coaches/players (χ^2 $p < .001$), possibly reflecting intensified attribution and scrutiny.

We argue that platformed fandom is a robust site of ambient affiliation and stance calibration: stance resources (pronouns, appraisal, referential choices) shift predictably with performance context, enacting solidarity, blame, celebration, and rivalry in ways that both reflect and shape communal identity. The study contributes (i) a large-scale Portuguese-language case to digital sociolinguistics, (ii) a replicable CADS–SFL workflow including pro-drop subject recovery for analyzing interpersonal meaning at scale, and (iii) evidence that team performance conditions interpersonal meaning in online communities.

#136

Neutralization and cue-weighting of sibilant contrasts in Parkinson's speech: a cross-corpus study

Firoz Ahmed, Ratree Wayland

To investigate which sibilant contrast place (/s–ʃ/) or voicing (/s–z/) is more vulnerable to phonological neutralization in PD patients' speech, this study aims to analyze matched read-speech samples from two PD patients' speech corpora (English - MDVR-KCL, Slovak - EWA-DB). In the Slovak EWA-DB, there are 1,649 participants, including those with Parkinson's disease, who engage in various speech tasks (vowel phonation, diadochokinesis, reading, naming, and picture description) for 30–45 minutes per session. The English MDVR-KCL dataset includes 37 participants (16 with Parkinson's) who completed reading and spontaneous dialogue tasks over 114 minutes, with an average segment length of 93.7 seconds. A Spanish /s/ replication (NeuroVoz) is planned as secondary evidence, if accessed. Audio will be normalized (mono 16 kHz, peak-normalized, 60 Hz high-pass filter), force-aligned and tokenized. Place contrasts will be measured using fricative centroid (M1), and voicing contrasts by percent periodic frames during frication and a low-band energy ratio. Neutralization will be indexed by reduced acoustic separation (ΔCentroid, ΔVoice), d'/Bhattacharyya distance, and diminished decodability in speaker-held-out classification (AUC). Mixed-effects models will test interaction between Group (PD vs. control) and Contrast (/s–ʃ/ vs. /s–z/), with position and vowel class as covariates, and corpus, speaker, and word as random effects. Drawing on Licensing-by-Cue and Dispersion, this study predicts greater contrast reduction in cue-weak contexts (codas; back/rounded vowels) and anticipate that voicing neutralizes earlier than place. Beyond empirical findings, the study contributes a transparent and replicable framework by comprising token lists, TextGrids, code, and feature tables for identifying which phonological dimensions are first compromised in PD speech.

References

Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. The Journal of the Acoustical Society of America, 108(3), 1252–1263. https://doi.org/10.1121/1.1288413

Mendes-Laureano, J., Gómez-García, J. A., Guerrero-López, A., Luque-Buzo, E., Arias-Londoño, J. D., Grandas-Pérez, F. J., & Godino-Llorente, J. I. (2024). NeuroVoz: a Castillian Spanish corpus of parkinsonian speech. Scientific Data, 11(1), Article 1367. https://doi.org/10.1038/s41597-024-04186-z

New Data Acquisition Findings Reported from King's College London (RADAR-Base: Open Source Mobile Health Platform for Collecting, Monitoring, and Analyzing Data Using Sensors, Wearables, and Mobile Devices). (2019). In Health & Medicine Week (p. 280). NewsRX LLC.

Rusko, M., Sabo, R., Trnka, M. et al (2024). Slovak database of speech affected by neurodegenerative diseases. Sci Data 11, 1320 . https://doi.org/10.1038/s41597-024-04171-6

#139

Lexical complexity as a lens on varied academic experiences in L1 and L2 postgraduate writing

Maha Al-Harthi

The language sciences have long aimed to understand how linguistic variation both reflects and shapes human experience. This study contributes to that goal by examining lexical complexity in postgraduate academic writing across first language (L1) and second language (L2) contexts. Variation is viewed not only as a marker of linguistic proficiency but also as a lens through which to view learners' positions within global knowledge production.

Drawing on two specialized corpora—the Corpus of Arab Proficient Users of English (CAPUE), comprising Saudi postgraduate dissertations, and the Corpus of English Native Speakers (CENS)—the study employs the Lexical Complexity Analyzer to assess 25 measures of lexical density, sophistication, and diversity. The findings reveal distinct patterns: although L1 and L2 writers attain similar levels of lexical density, L1 dissertations consistently exhibit higher lexical sophistication and more variation. In particular, L2 writers tend to use less frequent and more advanced lexical items less often and display less variation in verbs, nouns, adjectives, and adverbs. These results suggest that lexical complexity encodes broader experiential differences, including the challenges of producing high-stakes academic texts in a non-native language and negotiating disciplinary authority across linguistic boundaries.

The study emphasizes the significance of corpus-based methods in capturing the nuanced aspects of linguistic variation. By analyzing lexical features across extensive collections of authentic texts, corpus linguistics offers a powerful means of investigating the uneven development of L2 writing. At the same time, the findings invite a reconsideration of how "proficiency" is conceptualized. Rather than treating L2 differences solely as deficits relative to native-speaker norms, the results encourage a more context-sensitive perspective that acknowledges how lexical choices embody communicative strategies, cultural backgrounds, and academic paths.

Ultimately, this research shows how corpus linguistics links textual features with broader questions of academic participation. It reveals how lexical choices reflect learner experiences and institutional expectations, highlighting variation as a crucial aspect of multilingual academic practice. By doing so, it aligns with the conference theme by demonstrating that variation is not just noise but a reflection of the real lives of multilingual academic writers.

#141

Bootstrapping tupleised statistics to compare construction inventories: the case of Cantonese and Mandarin particle frames

Ryan Ka Yau Lai

Languages are often noted to be richer in certain grammatical constructions than others (e.g. Hindi's light verb constructions). Relatively little work, however, compares constructional inventories across language varieties using corpora. This talk examines particle frames, combinations of a final particle (FP) and a non-final particle (NFP) encoding discourse functions like focus and interrogativitiy, in Cantonese and Mandarin Chinese. Particle frames are traditionally considered characteristic of Cantonese but rare in Mandarin.

I extract FP and NFP tokens from the Hong Kong Cantonese Corpus (Luke & Wong 2015) and the Mandarin CallHome corpus (Liu et al. 2004). To determine which particle pairs are conventionalized constructions, I move beyond using a single statistic like pointwise mutual information (PMI) and adopt tupleisation (Gries 2023): Using multiple co-occurrence statistics measuring different aspects of co-occurrence. Specifically, I use PMI for bidirectional attraction and repulsion, Gries' (2022) KLD-based measure for unidirectional attraction, and entropy to measure particles' flexibility in 'choosing' particles to co-occur with. These measures are used to compute inventory size estimates by counting how many items exceed a set threshold (e.g. PMI > 3) for each variety.

The complexity and noisiness of corpus data, including differences in corpus size, long-tailed frequency distributions, statistical dependency within documents, and inherent bias in many statistical estimators, pose challenges for reliability and cross-corpus comparability. Like Gries (2023), I use a bootstrapping approach to overcome these challenges. I employ a Studentised bootstrap (Efron & Tibshirani 1998), which improves on the percentile bootstrap used in Gries (2023) by estimating how many standard errors a statistical estimate tends to diverge from the true value; this better accounts for the extreme bias in many of these estimators.

After bootstrapping, results are as follows. Using PMI > 3 as a criterion for mutual attraction, there are more Cantonese than Mandarin particle frames, though the difference is smaller than previously thought: Cantonese has about 50% more (268-299 at 95% confidence) than Mandarin (172-209). In both varieties, it is primarily the NFP that attracts the FP, rather than the other way around: using normalised KLD > .8 as a criterion, the NFP-to-FP direction boasts 94-109 pairs in Cantonese and 67-78 in Mandarin, while the FP-to-NFP direction has 4-8 pairs in Cantonese and 1 in Mandarin.

Other measures reveal a major functional distinction between Cantonese and Mandarin. Entropy measures show Cantonese NFPs are much more flexible than Mandarin NFPs as to what FPs they co-occur with, while Cantonese FPs may be less flexible than Mandarin in choosing NFPs. Moreover, 56-68 Cantonese NFPs and no Mandarin NFPs prefer appearing without an FP. These results suggests that while Mandarin has functionally very heavy NFPs and very light FPs, Cantonese has somewhat lighter NFPs and heavier FPs. Thus, FPs appear with a narrower range of NFPs, and NFPs would rather appear alone when no semantically compatible FP is available. This suggests that the key difference between Mandarin and Cantonese is not the number of particle frames, but how selective the particles are in co-occurring.

#150

Linking adverbials in academic writing: insights from L1 and L2 student corpora

Duong Nguyen

Academic writing has posed challenges to EAP learners from different proficiency and academic levels, regardless of their L1 background. One area that appears to be significantly challenging to these learners is cohesion, which could be achieved through five types of cohesive: reference, substitution, ellipsis, lexical cohesion, and conjunctions (Halliday and Hasan, 2013), or linking adverbials (Biber et al., 2021). Among these, linking adverbials (LAs)—due to their variation in form and meaning—appear particularly difficult for learners. Prior studies often benchmarked L2 usage of LAs against L1 groups, yet patterns of L1 usage have not fully been explored. Additionally, it is unclear whether academic level or disciplinary differences influence LA use in both L1 and L2 writing. To address these gaps, this study investigates the usage of linking adverbials (LAs) and their semantic types in academic essays of L1 and L2 English university students in the British Academic Written English (BAWE) corpus (1,220 texts, 3,222,628 words) across four academic levels (i.e., first-year, second-year, third-year undergraduate, master) and three disciplinary groups (i.e., Arts and Humanities, Social Sciences, Life and Physical Sciences).

AntConc 3.5.9 (Anthony, 2020) was used to extract linking adverbials (LAs) from L1 and L2 English learners' writings. The raw overall and individual frequencies of the linking adverbials (LAs) in the sub-corpora were computed and normed to 100,000 words. Chi-square tests were then performed on the normed frequencies of LAs to assess the significance of relationships between learner groups (L1/L2), academic levels, disciplinary groups, and LA semantic types. Additionally, observed and expected counts were used to calculate the Chi-square value for each cell in the contingency tables to identify which categories accounted for the differences, and qualitative analyses focused on the most frequent LAs (40 occurrences per 100,000 words).

Findings show that L2 writers employed more LAs in their academic essays (1,356 per 100,000 words) than their L1 counterparts (1,154 per 100,000 words), although both groups relied predominantly on a limited subset of this linguistic feature in academic writing, including also, however, therefore, then, thus, for example. Another important finding is that while disciplinary variations in LA usage were observed for both L1 and L2, there was no significant difference in the use of LAs across academic levels. Notably, the findings also indicate that while L1 learners used adversative LAs (e.g., yet) significantly more frequently than L2 counterparts, L2 learners employed significantly more additive LAs (e.g., moreover) and causal LAs (e.g., thus) than their L1 peers. The findings offer insights into the differential cohesive strategies employed by L1 and L2 writers and have important implications for targeted academic writing instruction aimed at enhancing essay cohesion for undergraduate students in diverse disciplinary contexts.

References

Anthony, L. (2020). AntConc (Version 3.5.9) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Biber, D., Johansson, S., Leech, G. N., Conrad, S., & Finegan, E. (2021). Grammar of Spoken and Written English. John Benjamins Publishing Company. https://doi.org/10.1075/z.232

Halliday, M. A. K., & Hasan, R. (2013). Cohesion in English. Routledge.

#151

Surveying native speakers' real time register use to design a representative corpus

Andrea Flinn

A need exists in the humanitarian aid sector for non-native speakers of Levantine Arabic, which is spoken in Jordan, Lebanon, Palestine, and Syria. However, corpora that represent Levantine Arabic remain limited, as do pedagogical materials for these dialects. The Levantine Arabic corpora which do exist are narrowly focused and are rarely based on a careful domain description (Althobaiti, 2020). However, conducting a careful domain description is the first step in principled corpus construction (Egbert et al., 2022).

The purpose of this study is to describe the proportions of registers within Levantine Arabic in order to guide the construction of a 2.5+ million word corpus which represents this variety of Arabic, so that ultimately corpus-based pedagogical materials can better represent these dialects. The registers used were investigated by conducting a Parameters of Language Use Survey (PLUS), which was designed by Hashimoto (2020) to describe the language use of a given population. When conducting the survey, I had 16 native speakers log their language use over a four-hour period. Participants recorded each language use in a paper logbook in real time. Later, they completed a Qualtrics survey about each language use, sorting it into one of the registers listed. I also used a mobile application tracker to record participants' phone use during the four hours. The tracker monitored the number of minutes participants spent on each mobile application. Additionally, I conducted two informal observations. First, I accompanied one participant during data collection and observed their logging behavior, its accuracy, and the problems that arose during the data collection process. Second, I observed the activities of non-participants in three public places including a public library, a coffee shop, and the walkways of a college campus, to gain a broader perspective on language and register use in Levantine Arabic. Lastly, participants completed a final survey in which they reported on the quality of their data collection.

Application tracker data revealed inconsistencies in the data of one participant, who was then eliminated from the study. Final surveys and Observation 1 indicated that data collection was largely accurate, and Observation 2 showed that non-participants' register use did not differ widely from that of participants. Findings featured the registers used (e.g. Conversations, music lyrics, audio/video sharing, and written messaging and social media interactions) and the total proportion of time spent using each register. As expected for a traditionally oral variety of Arabic, much language use consisted of conversation (61.1%) and song lyrics (9.2%). Another 16.4% consisted of digital language use, most of which was written, reflecting a noteworthy change precipitated by the advent of Web 2.0. These findings capture unexpected variation in language experience, specifically register use, within Levantine Arabic. Results can be used to compare varieties of Arabic including Modern Standard Arabic, and findings have informed the design of a 2.5+ million word corpus, as well as a 600-item pedagogical word list.

#155

Disciplinary register variation across the 'ages':
from undergraduate writer to researcher-in-training to disciplinary expert

Bethany Gray, Febriana Lestari, Duong Nguyen, Kimberly Becker

Given the prominent role and multiple purposes of writing in academic contexts, it is not surprising that research on academic writing is extensive and varied. Such research has documented systematic variation between disciplines, sub-registers (e.g., research articles vs. textbooks vs. essays), and 'novice' and 'expert' writers. One challenge has been unentangling the issues of writing development and register variation. That is, we know that writers at different levels of the academy exhibit varied patterns of language use, but we also know that they need to produce fundamentally different types of texts. While it may be possible to tease apart these factors by controlling for either novice/expert status or register, this approach would lack ecological validity. In this study, we prioritize ecological validity (describing the texts that writers produce at different academic stages) through a TxtLx register-based approach (Biber & Conrad, 2019; Biber et al., 2021) that systematically accounts for both situational and linguistic variation. We focus on disciplinary variation across across four academic stages:

- undergraduate students writing for their discipline-specific coursework,
- early-career graduate students (researchers-in-training) producing texts for their graduate coursework,
- doctoral students completing scholarly dissertations (the culminating text which transitions them to 'expert' status), and
- professional academics publishing peer-reviewed scholarship for the purpose of research dissemination.

The study is based on a cross-sectional compilation of corpora in two disciplines (engineering and linguistics), with sub-corpora to represent a range of sub-registers, sampled from the real-life academic contexts relevant to each group of writers. Texts to represent undergraduate student writing (N=332) are drawn from MICUSP (Römer & O'Donnell, 2004) and BAWE (Nesi et al., 2004-) across the five most common registers in these corpora. Several sub-corpora from CorGrad (Becker, 2022) are used to represent writing produced by graduate students (N=927) in graduate-level courses, encompassing 7 specific registers. The Dissertation Register Corpus (DRC; Lestari, forthcoming) contains successfully defended doctoral dissertations (N=86) in engineering and applied linguistics, including four major dissertation types. Finally, published academic writing is represented by two sub-corpora of texts published in peer-reviewed academic journals (N=300), including both empirical research articles and evaluative texts (such as forums and commentaries).

Following the register tradition (Biber & Conrad, 2019; Biber et al., 2021), this study carries out both situational and linguistic analysis. The situational analysis is based on a synthesized framework to describe the key non-linguistic characteristics of the types of writing represented by the corpora (through analyses of the domain and of the texts in the corpus). The linguistic analysis describes patterns of variation in the use of lexico-grammatical features through additive multi-dimensional analysis (Berber Sardinha, 2021), based on the five dimensions of variation in disciplinary writing identified in Gray (2015). The results show systematic variation across disciplines, registers, and novice/expert writers, although not always in expected ways (e.g., dissertations show surprising variation when compared to research articles). This linguistic variation is interpreted with direct reference to the situational characteristics of the texts/registers, providing insight into the incremental variation observed both situationally and linguistically across these four stages of the academy.

#158

Uncovering cross-linguistic register variation with a shared MD model of English and Chinese

Shangyu Jiang

Cross-linguistic register comparisons are essential for testing whether register patterns reflect language-independent communicative functions or language-specific functions. Multi-Dimensional Analysis (MDA; Biber, 1988) has been highly productive within single languages and has informed claims about cross-linguistic universals (e.g., Biber, 1995) by comparing separately built MD models. However, such comparisons have limitations: selections of linguistic features differ by language, factor spaces are estimated independently, and factor scores are not directly commensurable. To address these constraints, this study presents a joint, cross-linguistic MDA fitted simultaneously to English and Mandarin Chinese, attempting to uncover register variation across both languages in a single, shared MD space.

The dataset consists of two Brown-family corpora with parallel register sampling: Crown (American English, ~1M words; Xu & Liang, 2013) and ToRCH2019 (Mainland China Chinese, ~1M words; Li et al., 2022). Thirty-nine functionally relevant linguistic features (largely borrowed from Biber, 1988) were operationalized in both English and Chinese and tagged in the dataset using spaCy's transformer models. The feature tags were first counted and normalized. The normalized counts were then z-scored within language before they were pooled for exploratory factor analysis (PAF; Promax). Diagnostics indicated strong factorability (KMO = .84; Bartlett $p < .001$) and a six-dimension solution was retained.

The resulting dimensions were preliminarily labeled as (D1) Clause-elaboration vs. phrasal description, (D2) Involved/interactive packaging vs. nominal packing, (D3) Copular/stance presentation vs. nominal compression, (D4) Modal/attitudinal description vs. human reference, (D5) Retrospective/passive reporting vs. parataxis, and (D6) Quantification/metrics. When compared on each dimension, most registers lined up similarly across English and Chinese, while some registers showed different patterns of functional variation across the two languages. For example, English learned/government writing sits on the retrospective/passive pole of D5, whereas Chinese learned/government writing leans toward parataxis; English press reportage is more clausal on D1, while the Chinese counterpart is slightly phrasal. These patterned similarities and differences indicate that a single shared MD model can both (i) more thoroughly uncover cross-linguistic universals of register variation and (ii) reveal language-specific register functional patterns. This cross-linguistic MDA provides a scalable method for comparing registers across multiple languages.

References

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

Biber, D. (1995). Dimensions of register variation: A cross-linguistic comparison. Cambridge University Press. https://doi.org/10.1017/CBO9780511519871

Li, J., Sun, M., & Xu, J. (2022). ToRCH2019 Corpus (ToRCH2019现代汉语平衡语料库): Texts of Recent Chinese corpus 2019 (A Brown family Chinese corpus of one million words) (Version 1.5) [Dataset]. http://corpus.bfsu.edu.cn/ToRCH2019.zip

Xu, J., & Liang, M. (2013). A tale of two C's: Comparing English varieties with Crown and CLOB (the 2009 Brown family corpora). ICAME Journal, 37, 175–184. http://icame.uib.no/ij37/Pages_175-184.pdf

#164

Corpus approaches to media, language, and the politics of recognition

Caroline Scheuer Neves

This paper addresses what can be described as a paradox of inclusivity in Brazilian online news discourse between 2012 and 2019. Drawing on the Brazilian subcorpus of the News on the Web (NOW) Corpus (Davies, 2018), the study examines how LGBTQIAPN+-related vocabulary and nonbinary forms circulate within reporting in Brazil. While debates around nonbinary and inclusive Portuguese have gained more attention in recent years, little empirical research has traced how these forms are actually taken up in large-scale public discourse. This project responds to that gap by documenting the progress and the limits of inclusivity in news media language.

The analysis combines frequency counts, collocational profiles, and concordance exploration to investigate patterns across seven years of news texts. Results reveal that broad umbrella terms such as LGBT and trans gained steady and increasing circulation, often embedded in contexts tied to legislation, political struggles, and health care. These forms stabilized as recognizable markers for representing gender and sexual diversity in public discourse. However, queer nonbinary identities and innovative linguistic resources, including pronouns (elu, ile) and morphological variants (-x, -e), remained almost entirely absent. When such terms did appear, they were typically confined to reports on activism or institutional statements, and rarely entered the domain of everyday news reporting.

This divergence highlights a paradox: inclusivity at the umbrella level does not automatically translate into inclusivity at the queer nonbinary level. While the media expanded its lexicon to acknowledge LGBTQIAPN+ communities in broad terms, it continued to marginalize and exclude linguistic innovations that explicitly index nonbinary subjectivities. Such selective uptake reinforces hierarchies within the LGBTQIAPN+ representation, privileging identities that are more institutionally recognized while leaving others at the margins of public discourse.

The uneven distribution of these forms also reflects broader sociocultural and political conditions in Brazil. Inclusive Portuguese has been the subject of proposals from activist groups and educational institutions, but its implementation has been fragmented and contested. Media outlets, functioning as key gatekeepers of linguistic legitimacy, demonstrate caution in adopting nonbinary innovations, a tendency likely shaped by political polarization, institutional inertia, and concerns over audience reception. In this way, the paradox of inclusivity is not only linguistic but also ideological: it indexes tensions between symbolic recognition, institutional authority, and advocacy.

By documenting these dynamics, the study highlights the significance of corpus-based approaches in understanding how inclusivity is negotiated in real-world discourse. The findings contribute to broader conversations about the role of language in mediating social recognition, showing how news discourse can advance and constrain the possibilities of inclusive languaging. Ultimately, the paper argues that media texts do more than reflect existing social attitudes; they actively participate in shaping the symbolic boundaries of belonging.

Poster Abstracts

#2

Contrastive analysis of mood and modality in English and Naija

Abiola Iyiade

Several studies have observed that the main obstacle to second language learning is the interference of learner's native language system. Nigerian Pidgin has attracted scholarly attention in the last decade, especially in Nigeria's national language discussion (Eligbe 1995). Most of the past studies have examined various aspects of Nigerian Pidgin from the perspectives of sociolinguistics, literary writing, profiling, cultural influence etc., little attention (if any) has been given to features of mood and modality as aspects of syntax in English and Nigerian Pidgin. This study, therefore, undertakes a contrastive analysis of mood and modality in English and Nigerian Pidgin. The objectives, among others, include describing the nature of mood and modality in both languages and identifying the differences between mood and modality systems in the two languages as a means of predicting what learning problems Nigerian Pidgin learners may encounter in learning English in the L2 environment.

The theoretical pivot for analysis in this study is the systemic function grammar (SFG) where mood and modality are captured. The sub- theory selected is interpersonal metafunction. The data for the study consist of sentences from the NaijalexElab corpus and Naijionary database. The selection is based on the relevance of the sentences to the study. The data were collected from Sapele-Warri-Ughelli in Delta State where people use the language as native speakers. The data were transcribed using ELAN annotated for macrosyntax. The sentences are subjected to structural analysis in order to account for the focus of the research.

The two languages exemplify the features of mood and modality. The two languages also display the presence of mood types and modal operators. Polar question in English has subject-operator inversion e.g (will there be another attack ?) while polar question in Nigerian Pidgin has a declarative sentence signalling polarity by changing falling intonation contour to rising contour at the final.syllable. For example (you don baf ?). Tag questions are formed in different ways in English especially by using auxiliary verb from the sentence e.g (I cannot believe the report , can I ?) while Nigerian Pidgin has few ways in expressing tag question e.g (abi). English modal operators are tense marked e.g (will- would). Nigerian Pidgin modal opeartors are not marked for tense.

This study submits that the areas of similarity will enhance learning of English while the areas of divergence would pose problems for the Nigerian Pidgin learners of the English language.

#9

A corpus-based contrastive study on the behavioral profile of adverbs of degree in Mainland and Taiwan Chinese

Yan Xiao

Adverbs of degree are the most different adverbs between Mainland and Taiwan Chinese. Exploring the similarities and differences of adverbs of degree on both sides of the strait based on the corpus not only helps to reveal the subtle differences language variation, but also helps to broaden the path for computer-assisted compilation of various language dictionaries, and broadens the application scope of corpus in second language acquisition and research. This paper compares 33 adverbs of degree in Mainland and Taiwan Chinese by using BP (behavior profile) analysis based on a self-built corpus, and finds that: (1) The use of adverbs of degree in Mainland and Taiwan Chinese are significantly related to the semantic prosody, part of speech, and syntactic components of the words they modify. (2) Adverbs of degree in the Mainland corpus modify more positive semantic prosody than those in Taiwan; Adverbs of degree in Chinese corpus mostly modify verbs, while adverbs of degree in Taiwanese corpus mostly modify adjectives. (3) High-magnitude and low-magnitude adverbs of degree are more used and extreme and low-magnitude adverbs of degree are avoided in both corpora while Taiwan uses more extreme adverbs of degree than Mainland. This article further points that the similarities and differences between adverbs of degree in Mainland China and Taiwan are mainly influenced by language style norms, cultural system norms, and translational shining through.

#16

Overcoming database and software challenges in corpus linguistics research in developing countries: a case study of Africa

Solange Swiri Tumasang

Corpus linguistics has emerged as a powerful methodology for analyzing language patterns and structures, yet its application remains limited among scholars and researchers in developing countries, particularly in Africa. The lack of access to robust databases and specialized software hinders the growth of corpus linguistics research in the region. This study investigates the database and software challenges faced by researchers in Africa when conducting corpus linguistics research. Through a mixed-methods approach, combining surveys, interviews, and case studies, we examine the specific difficulties encountered by African researchers in accessing and utilizing corpus analysis software and databases. A total of 150 researchers from various African institutions participated in the online survey, while 30 researchers were selected for in-depth interviews and 5 case studies were conducted to gain a deeper understanding of the challenges faced. Our survey instrument consisted of 25 questions, covering topics such as access to corpus analysis software, database availability, and technical support. The interview protocol was designed to gather more nuanced insights into the challenges faced by researchers and potential solutions. Our findings reveal that limited access to proprietary software, inadequate infrastructure, and lack of technical support are significant obstacles to corpus linguistics research in Africa. Specifically, 80% of survey respondents reported difficulty accessing proprietary software, while 70% cited inadequate infrastructure as a major challenge. Furthermore, our case studies highlight the resourcefulness and creativity of African researchers in developing context-specific solutions to these challenges. We also identify potential solutions, including the development of open-source corpus analysis tools and context-specific databases. The study's outcomes have implications for researchers, institutions, and policymakers seeking to promote the growth of corpus linguistics research in Africa and other developing regions. By shedding light on the challenges and opportunities in corpus linguistics research in Africa, this study contributes to the development of more effective strategies for supporting linguistic research and language development in the region. The findings of this study can inform the development of corpus linguistics curricula, research capacity-building programs, and infrastructure investments in Africa. Ultimately, this study aims to promote the advancement of corpus linguistics research in Africa and contribute to the global body of knowledge in the field.

#20

Temporal fluency and lexical diversity:
a comparison of their predictive power on the Spanish OPIc

Alan Brown, Greg Thompson, Troy Cox, Earl Brown

Researchers frequently use the acronym CALF (Complexity, Accuracy, Lexis, Fluency) when referring to various characteristics of L2 speech. Of the four components of the acronym, Lexis and Fluency have become the most amenable to large-scale, objective, and automated analyses. Indeed, as measures of temporal fluency go up so too do proficiency scores (Cox et al., 2023). Overall speech rate and mean length of utterance seem to be the best predictors of global proficiency, with the former measuring how much language is produced over time and the latter measuring how much language is produced between pauses (De Jong & Bosker, 2013; De Jong, et al., 2015). When raters were asked to assign a proficiency level to speech samples using a holistic approach, the most powerful predictor was speech rate (Iwashita, et al., 2008) with lexical density as the second most predictive feature. Thus, DeJong (2016) has argued that fluency "is among the most important aspects of speech that makes an L2 speaking performance successful" (p. 207). Yet, greater lexical density, or lexical diversity (LD), has also been shown to maintain a strong, positive relationship with L2 proficiency, across diverse languages (Kyle et al., 2021; Lee, 2019; Treffers-Daller et al., 2018; Woods et al., 2023), including L2 Spanish (Consolini & Kyle, 2024; Fernández-Mira et al., 2021; Tracy-Ventura et al., 2021).

The current study examines the relationship between temporal fluency, LD, and scores on the ACTFL OPIc (Oral Proficiency Interview by computer), whose ratings seem to align with the OPI (Thompson, et al., 2016). The L2 Spanish corpus used for the analysis comes from 65 Spanish learners who orally responded to between 13 and 17 prompts each with a maximum recording time of two minutes per prompt. Student responses were transcribed using the medium model of the open-source automated speech recognition software Whisper after it was determined that the software achieved, on average, a 93% word accuracy rate after 20 transcribed responses were compared to human transcriptions—quite high for L2 speech.

A total of 1,019 responses were transcribed, totaling 162,215 words and 1,675 minutes (~28 hours) of audio with responses ranging from 42 to 383 words. Lexical diversity was measured using the Moving Average Type to Token Ratio (MATTR), the Measure of Textual Lexical Diversity-wrap around (MTLD-wrap), and the Hypergeometric Distribution D (HD-D) while articulation rate as number of syllables per second of speech, mean length of utterance as number of syllables per utterance, and silent pause ratio as number of silent pauses divided by the total time were measured as indices of temporal fluency using Praat. A multiple regression analysis was conducted with the MCMCglmm R package (Hadfield, 2010). Each learner's official ACTFL oral proficiency rating served as the dependent, or outcome, variable, which was treated as an ordinal variable. Participants' ratings ranged from Intermediate Low to Superior. As proficiency level increased, so did text length, all three LD measures, articulation rate, and mean length of utterance while silent pause ratio was unaffected by proficiency level in these data.

#23

Noun phrase development across registers in the first two years of Chinese learning

Yilei Li

Writing and speech are characterized by different syntactic features, which pose challenges for learners to differentiate ways of increasing syntactic complexity across registers. Among the syntactic complexity measures, the noun phrase (NP) has been found typical in formal writing and related to writing scores in second languages (L2s). However, little is known about the development of NP complexity in L2 Chinese, including its register differences from the learners' practices. Focusing on two specific types of NP in Chinese, this study uses written and spoken data from Chinese learners in The Multilingual Academic Corpus of Assignments - Writing and Speech to explore whether learners employ more complex NPs in writing rather than speaking, as they go through stages of learning. It turned out that NP complexity in writing increases from the beginning to the intermediate level classes, while it does not develop much in speaking, which suggests learners' growing register awareness in terms of syntactic complexity. Meanwhile, the second and third semesters of learning are found to be the crucial period when learners struggle to manage NP complexity. Based on the preliminary findings, future research could explore the development of other types of NP and typical syntactic features in spoken Chinese.

#28

Can AI sound like a teacher? Corpus evidence from human and simulated classrooms

Marilisa Shimazumi, Tony Berber Sardinha

This study investigates how English language teachers perceive and evaluate classroom discourse produced by humans and by LLMs simulating teaching roles. With the expanding use of generative AI in educational contexts (ranging from material preparation to simulated instruction), the question is no longer whether AI can be used in teaching, but how convincingly it performs in relation to professional expectations. Specifically, this study addresses how educators interpret the pedagogical quality of AI-simulated instruction and whether such discourse can convincingly pass as human-produced. The corpus consisted of 19,950 online class segments (28,054,701 words). These were GPT 4o-generated teacher dialogs that mirrored topics and teacher-student interaction sequences from human-taught classes, with the AI prompted to assume one of 56 different personas, which varied across affective traits, demographic profiles (such as age, ethnicity, gender, and sexual identity), pedagogical orientation, and other characteristics. From this corpus, which was analyzed using Lexical Multi-Dimensional Analysis (Berber Sardinha & Fitzsimmons-Doolan, 2025), five dimensions emerged: Dimension 1 distinguishes an organized, plan-driven style from a performance-focused one. Dimension 2 contrasts cooperative scaffolding with authoritative confrontation. Dimension 3 opposes procedural control to informal rapport-building. Dimension 4 separates evaluative pressure from reflective, aesthetic engagement. Dimension 5 differentiates informal facilitation from formal, deductive exposition. Based on these dimensions, 20 class segments were selected, comprising both human and AI versions; these were rated by 31 EFL teachers using TRACE (Teacher Rating for Assessing Classroom Effectiveness), a protocol specifically designed for this study to assess instructional quality, and to indicate whether they believed each excerpt was produced by a human or by AI. The results showed teachers correctly identified the source in the majority of cases, with higher accuracy for human-produced texts than AI-simulated ones. However, there was considerable variation across raters, and a surprising pattern emerged: greater teaching experience was linked to lower accuracy in source identification. This suggests that more experienced educators may have been less suspicious of atypical styles, possibly because they have encountered a wider range of teaching personalities and techniques throughout their careers. Despite the relatively high identification accuracy, teachers were more likely to assign lower quality ratings to excerpts they believed were AI-generated, even when those excerpts were, in fact, human. This phenomenon, referred to as "AI stigma" or "algorithm aversion," appears to influence evaluations regardless of actual source. The stigma may stem from perceptions that authentic teaching requires personal experience, empathy, and a capacity for care, which are qualities not associated with machine-generated discourse. A success case involved an AI simulation that was frequently mistaken for human due to its use of interactional strategies associated with real-time spontaneity, such as hesitations, hedging, informal phrasing, and adaptive behavior. These characteristics reduced social distance and encouraged participation, reinforcing the illusion of human authorship. The findings suggest that although AI can produce linguistically competent and pedagogically plausible discourse, instructional quality judgment is strongly associated with perceptions of authenticity.

Reference

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

#29

Rewriting immigration: discursive shifts in AI-generated news

Tony Berber Sardinha, Marilisa Shimazumi

News media serve as a powerful channel for circulating public discourse, influencing how events, identities, and ideologies are conceptualized. With the increasing use of AI in journalism, the question arises of the underlying representations encoded in machine-generated texts. These representations often remain implicit, making it difficult to assess how generative AI may influence public conversation around complex issues such as immigration. To investigate this, we developed a corpus-based simulation of news reporting on immigration. We began by extracting 110,873 immigration-related articles published in the United States and the United Kingdom between 2016 and 2020 from the NOW (News on the Web) corpus (Davies, 2016). Articles were filtered to ensure immigration was the primary theme, and a random sample of 300 texts per quarter was drawn. From this pool, 1,200 texts were selected based on political orientation (400 left-leaning, 400 right-leaning, and 400 without clear ideological marking). Each human-written article was summarized and passed to GPT via API. The model was instructed to generate three new articles for each summary, adopting left-wing, right-wing, and non-partisan perspectives. This process yielded 3,600 synthetic articles, resulting in a combined corpus of 4,800 texts and approximately 7.6 million words. We applied Lexical Multi-Dimensional Analysis (LMDA; Berber Sardinha & Fitzsimmons-Doolan, 2025) to identify dimensions of variation in the corpus. LMDA uses factor analysis to detect recurring lexical groupings across a large set of texts. These groupings are interpreted as discourse poles that reveal recurring communicative patterns in how immigration is represented. The method differs from dictionary-based approaches by focusing on emergent lexical configurations rather than predefined categories. Seven dimensions were identified, each distinguishing two opposing discourse poles. For instance, Dimension 1 contrasts representations of immigration as an existential and symbolic threat with portrayals grounded in the everyday realities of housing, labor, and social services. Dimension 2 opposes legal and procedural framings to geographic and geopolitical ones. Dimension 3 contrasts progressive moral defense of immigration with representations framing immigration as a threat to morality and public safety. A total of 14 discourse types emerged from these contrasts. Dimension scores were computed for each article, and ANOVA tests were used to compare the distribution of discourse types across human and AI-generated texts. Results showed that AI texts more frequently adopted progressive representations and did so more consistently than human-written counterparts. In some cases, AI-generated right-wing articles appeared to soften positions typically found in human-authored texts. Nonetheless, AI occasionally produced high scores on discourses that framed immigration as a threat to national sovereignty or as criminalized mobility. These findings complicate the view that AI merely reproduces conservative or centrist bias. Instead, the results point to a reconfiguration of discourse patterns, influenced by prompt design and model behavior. As generative AI becomes more widely used in journalism, its capacity to reconfigure the public understanding of sensitive topics (subtly or overtly) merits close attention.

References

Berber Sardinha, T., & Fitzsimmons-Doolan, S. (2025). Lexical multidimensional analysis. Cambridge University Press.

Davies, M. (2016). Corpus of news on the web (NOW). https://www.english-corpora.org/now

#36

A negotiated communion in parliamentary debates:
concerns and representations of hospice-related objects in Hansard

Yan Cheng

Parliamentary discourse has always been considered institutionalized multi-party confrontational political discourses. However, what has often been overlooked is that parliamentarians engage not only in confrontational, but also in collaborative work, establishing non-partisan ideological commitments and thus creating a special sense of communion. This study combines structural topic modeling and discourse analysis to analyse 1050 debates about 'hospice' from 1928 to 2025. The distribution of parliamentary documents related to hospice care reveals four distinct periods of heightened attention, three of which occurred under Conservative governments and one during Labour's administration. While initial discussions began during Conservative rule in the 1980s, the topic gained notable momentum under the Labour government (1997–2010), with a sharp rise in document frequency starting in the early 2000s. Continued growth was observed during the Conservative–Liberal Democrat coalition (2010–2015), followed by a marked surge in attention under the Conservative government post-2015. Notably, 2024 and 2025 witnessed unprecedented peaks, with 72 and 91 documents respectively, reflecting a significant increase in political and policy engagement with hospice care in recent years.

A comparison of the top 100 collocates of hospice across four political periods reveals a dynamic evolution in the language and framing of hospice care in UK parliamentary discourse. As illustrated in the stacked bar chart, the earliest period (P1: 1979–1996) contains 38 unique collocates, reflecting a discourse rooted in emotional appeal, local voluntary efforts, and foundational narratives—evident in words such as marvellous, terminally, and voluntary. The Labour government era (P2: 1997–2005) maintained some of these expressive tones while introducing terms related to ethical debates and social advocacy, such as euthanasia and petitioners, marking a shift toward legislative and rights-based concerns. The coalition period (P3: 2006–2014) exhibited a more professionalized and institutional tone, incorporating references to international models (oregon) and care systems (nightingale, farmhouse). Most notably, the most recent phase (P4: 2015–2025) is distinguished by the highest number of unique collocates, signaling a clear lexical shift toward technocratic and policy-driven language. This progression from emotionally resonant language to strategic policy terminology highlights the increasing institutionalization and politicization of hospice care over time.

Using a structural topic model with year and house as covariates, five distinct topics were identified. Among these, healthcare system & voluntary care, end-of-life ethics & euthanasia were found to be significantly more prominent in debates within the house of commons, whereas hospice & palliative care funding, political discourse on public health & aids, and charity, tax policy & public support were more extensively discussed in the house of lords. With regard to temporal variation in debate intensity, references to euthanasia have demonstrated a notable upward trend in recent years, while interest in healthcare system & voluntary cancer care has markedly declined. A more nuanced discourse analysis further indicates that, in both houses, individuals and groups associated with hospice are frequently constructed through discursive representations such as dedication, contribution, tremendous work, and compassion. Members of parliament also routinely invoke their personal experiences and involvement in the field as discursive strategies to legitimize their stances.

#37

Lexical appropriateness of the Korean College Scholastic Ability Test (CSAT) English: a comparative analysis with English Corpora

Nena Choi

South Korea's College Scholastic Ability Test (CSAT) remains the nation's clear high-stakes examination, linking university admission to future job-market outcomes. Sharply increasing private-tutoring costs have made the government to replace the English section's norm-referenced scoring with a criterion-referenced format, yet the lexical validity of the new test has not been examined. This study therefore profiles the lexical coverage of CSAT reading passages against three benchmark lists—the Ministry of Education's 3 000-word list (MOE-3k; 2022), the frequency bands of the Corpus of Contemporary American English (COCA; Davies, 2024a) and COCA Academic Word List (COCA-AWL; Davies, 2024b)—using AntWordProfiler 2.2.1 (Anthony, 2024). Results are interpreted through Nation's (2013) comprehension criteria (95 % minimally assisted; 98 % independent). The findings indicate that the MOE-3k list alone falls well below the 95 % threshold; even after adding high-frequency COCA bands, substantial lexical gaps remain. These gaps undermine the intended equity benefits of criterion-referenced scoring. According to the finding, this research suggests implications for policymakers and educators, including vocabulary-list refinement and syllabus alignment, to ensure that CSAT reading tasks match curricular goals while reducing dependence on private tutoring.

#45

Zooming in and out on the learner lexicon: a CDST approach to L2 Turkish vocabulary development

Yigit Savuran, Stefanie Wulff

While learner corpus research has been invaluable for identifying group-level developmental patterns, it often overlooks the variable and non-linear trajectories of individual learners. Adopting a Complex Dynamic Systems Theory (CDST) perspective (Larsen-Freeman & Hiver, 2025), this study utilizes the Turkish Learner Corpus (TURLEC) to demonstrate how both group trends and idiosyncratic learner paths can be analyzed within a single corpus to provide a richer understanding of L2 lexical development.

This paper presents a contrastive case study of carefully selected lemmas representing two distinct usage patterns. First, we analyze two "heavy hitter" verbs (olmak, 'to be'; etmek, 'to do'), characterized by high frequency and wide dispersion, tracing their collocational profiles to map the developmental patterns across learners. Second, we contrast these with two "idiosyncratic" lemmas (üvey, 'step-'; hırsız, 'thief'), characterized by high repetition but extremely low range (used by two learners).

Our analysis shows that while the heavy hitter verbs reveal systematic, group-level patterns, the idiosyncratic words highlight the crucial role of individual variability. Their usage can be seen as personal phase shifts into new lexical states, driven by life events rather than a linear curriculum. By juxtaposing these "zoomed-out" and "zoomed-in" perspectives, this study provides empirical evidence for core CDST tenets. We argue for a methodological approach that embraces both systematic patterning and individual variability to fully account for the complex process of language learning.

References

Larsen-Freeman, D. & Hiver, P (2025). Complex dynamic systems theory. In VanPatten, B., Keating, G. D. & Wulff, S. (Eds.), Theories in second language acquisition (4rd, 217–244). Routledge. DOI: 10.4324/9781003491118-9

#53

A corpus analysis of semantic drift and pejoration in English

Chad Hammock

As part of a larger project on semantic drift and pejoration, the current study aims to gauge the efficacy of corpus analysis in determining semantic drift. Anecdotally, word meanings and their cultural acceptability shift over time as, per Schellenberg (1996), younger generations shift towards an increased usage in taboo speech. While this seems to be common knowledge, there is a lack of corpus informed research on the topic. The current study proposes a data-centric approach which could produce results that are both reliable and robust.

The current study will utilize corpus tools to find and analyze semantic drift around English words in popular usage, both contemporarily and in the past, in an attempt to not only describe the semantic drift that has occurred but to determine the impetus of the change in meaning and cultural acceptability in order to predict future changes. The project will look at changes in word meaning over time by looking at both inoffensive and offensive language as demonstrated in corpus data through frequency and, where available, speaker data. This approach will allow for authentic results that can be analyzed with the benefit of multiple examples across longer time frames.

#57

LC-meta: a core metadata schema for L2 data documentation

Jennifer-Carmen Frey, Larissa Goulart da Silva, Alexander König, Hubert Naets, Egon Stemle, Magali Paquot

The main objective of this poster is to introduce a core metadata schema for learner corpora, developed through extensive collaboration between learner corpus compilers at the Centre for English Corpus Linguistics (UCLouvain, Belgium) and Eurac Research (Bolzano, Italy), and a research data infrastructure expert and member of CLARIN's metadata taskforce.

The project stems from the recognition that one area that would benefit significantly from standardization is L2 data description, which includes metadata at the level of the dataset as a whole and metadata used to describe the individual learners and task types/registers the corpus is meant to represent. There are a number of reasons why this is important. First, standardized and well-structured metadata increases the findability and usability of existing learner corpora. Second, it should enhance the comparability of datasets and comparability of L2 studies, provided researchers agree on a common set of definitions. Extensive metadata that follow - at best - a standardized vocabulary, and have a strong focus on findability, accessibility, interoperability, and reusability (FAIR) are an essential aspect of FAIR research data (Wilkinson et al. 2016).

In continuation of Granger and Paquot (2017), our proposed metadata schema is divided into a number of different sections for Corpus metadata (itself divided into Administrative metadata (e.g. authors or license) and Corpus design metadata (e.g. date and place of collection or type of task)), Text metadata (fine-grained per-text information), Learner metadata (details about the learners, e.g. age, languages spoken), Annotation metadata (e.g. details about manual or automatic annotation), Annotator metadata (e.g. professional and language background), Transcriber metadata (e.g. native language or language repertoire) and Situational and Task metadata (e.g. instructions, time constraints). While basic information about learners (authors) and language samples (texts) are typically found as part of metadata associated with a learner corpus, other aspects such as those related to the annotation or transcription procedure or the specificities of a task are often found elsewhere (e.g. corpus manual) or are just absent from currently available learner corpora.

A first version of the core metadata schema was tested on a range of learner corpora representing a variety of learner profiles and language samples (Paquot et al., 2023). It was presented at several conferences (e.g. LCR2022, EUROSLA2023), as part of an extensive feedback collection phase. Additional feedback was gathered via mailing lists. Based on the comments received, we substantially revised the initial proposal and released LC-meta version 2 in 2024 (Paquot et al., 2024a; 2024b). In 2025, a metadata working group was established under the aegis of the Leaner Corpus Association, in collaboration with the CLARIN K-centre for Learner Corpora. Its mission is to further develop, maintain, and disseminate LC-meta. Current work in progress includes: (1) designing a learner questionnaire to assist corpus compilers in collecting learner-related information, and (2) translating the existing schema into a machine-readable format. This format can serve as a foundation for implementation in corpus portals such as the CLARIN Virtual Language Observatory and as a resource for the future development of a learner corpus compilation infrastructure.

#59

Developing a corpus of reading errors in English early child language

Madison Rose, Michael Bennie, Valeria Pagliai, Walter Leite, Zoey Liu

Prior research has shown that production error patterns in early child language can inform characterizations of children's grammatical development (Smith 1933), provide insights into language acquisition theories (Locke 1980), and guide the advancement of language technologies designed for children (Booth et al., 2020; Jain et al., 2024). These findings can be further leveraged to design automated assessments critical for children's educational support and intervention (Paige 2020). However, existing studies on child production errors have been constrained by the sample sizes investigated (Stemberger 1989), as well as restricted public access to their datasets (Smith 1933). These limitations raise questions about the generalizability of previous findings, calling for larger, more accessible corpora with broader coverage of children's production errors.

We report ongoing development of a corpus of English reading errors produced by 63 children in elementary school, focusing on production errors occurring during naturalistic reading tasks. The corpus consists of 321 recordings (~9hrs), totaling 3,025 utterances. Each recording is a child's reading of a story tailored to their reading levels and learning differences. Data annotations are carried out by eleven undergraduate and graduate annotators with prior training in phonetics.

The annotation process involves utterance segmentation and error classification at both word and sentence levels. To provide a consistent framework for annotations, annotation guidelines are developed and refined throughout the annotation process. These guidelines detail operational instructions for the annotation interface, the annotation scheme, transcription conventions, and unusual case specifications. Compared to existing corpora of child reading errors (e.g., the CMU Kids Corpus by Eskenazi et al., 1997, which is not freely accessible), our annotation guidelines offer more details and clarity. We characterize 13 distinct error categories to cover different linguistic levels, including contraction/shortening, orthographic, phonological, grammatical, structural, visual tracking, run-on production, disfluency, self-response, unintelligible, and other. Each error category contains multiple sub-labels to accurately capture an error. For example, phonological errors are distinguished by whether they involve vowels or consonants and by the operation affecting them, such as insertion, substitution, or omission. Multiple error categories can be selected at once to investigate error co-occurrences.

Average story recordings lasted 1m31s, ranging from 0m07s to 5m23s and did not vary significantly by grade level ($p>0.10$). However, the speech rate of children in first grade is lower, averaging 68.4 words per minute (WPM), compared to 96.1 WPM for second grade and 96.8 for higher grades. Sentence-level annotations found 24 repeated and 208 run-on sentences. Word-level annotations show 63.7% of all words to be correctly produced. The most frequent errors were phonological (33.08%) and disfluency errors (30.09%). Average reading fluency measured as words correct per minute (WCPM) increased with higher grade levels (1st Grade: 51.7%; 2nd: 70.7%; 3rd and higher: 76.1%), but considerable variation exists within each grade level.

Upon completion of corpus development, we will release our data and annotations free of charge through the Linguistic Data Consortium. This corpus will support future research in automatic speech recognition, language acquisition theory, and speech pedagogy as an openly available resource for analyzing children's reading errors.

#61

Lexical multidimensional analysis of art discourse:
capturing human experience in the language of Sally Mann's photography

Yara Maria De Toledo Dias Romeiro

The visual arts are central to cultural life, yet the linguistic dimension through which they are mediated remains underexplored in language sciences. Texts that accompany, interpret, and circulate around works of art—whether in wall labels, memoirs, critical reviews, or exhibition catalogues—play a crucial role in framing the viewer's experience. They describe, evaluate, and theorize about artworks, thereby shaping both public reception and scholarly understanding. This research addresses this gap by applying Lexical Multidimensional Analysis (Lexical MDA) to the discourse surrounding the photography of Sally Mann, a major contemporary American artist whose work engages themes of family, mortality, intimacy, and Southern landscapes.A specialized corpus was compiled, consisting of 555 texts and 764,776 words spanning more than four decades. The corpus integrates twelve distinct registers, including books authored by the artist, exhibition wall texts, memoirs, press articles, and critical reviews, sourced from Sally Mann's library and public repositories. All texts were cleaned, lemmatized, and part-of-speech tagged. A total of 377 lexical variables were extracted and submitted to factor analysis, which revealed seven interpretable dimensions of discourse. These lexical dimensions capture the ways in which language constructs meaning around Mann's photography, highlighting recurrent discursive patterns that go beyond individual genres or registers. The dimensions identified reflect both the thematic concerns of Mann's oeuvre and the varied perspectives of those who write about it. They include: (1) Non-Specialized Opinion and Judgment, characterized by evaluative and subjective vocabulary; (2) Southern Landscapes / Dark Spirit / Photographic Process, revealing the interplay of geography, history, and technique; (3) Mortality in Exhibition, emerging from texts on her husband's illness and its photographic documentation; (4) Accounts of Daily Life, grounded in personal and anecdotal narratives; (5) Intimacy in Exhibition, which foregrounds debates around family photography and the ethics of representation; (6) The Photographer's Memoir, linking autobiographical discourse to artistic identity; and (7) Specialized Criticism and Artistic Appreciation, which engages in technical and symbolic analysis of Mann's work. Collectively, these dimensions trace how verbal language situates photography in cultural, ethical, and historical contexts. By investigating not the images themselves but the discourses that narrate and interpret them, this study demonstrates how Lexical MDA can systematically uncover patterns of variation in language use that reflect broader human experiences, such as grappling with death, negotiating intimacy, or situating art within cultural memory. The findings underscore the role of language as an indispensable mediator of visual art, revealing how words frame perception, generate controversy, and sustain artistic legacies over time. This research thus contributes to bridging corpus-based linguistic methodologies with visual arts studies, expanding the scope of language sciences to domains traditionally considered outside its reach. It illustrates how empirical, quantitative analyses of language capture the diversity of human engagement with art, and, more broadly, how discourse about images encapsulates the intersection of aesthetics, culture, and lived experience.

#70

Comparing computer-based and paper-based DDL for vocabulary learning

Dilay Candan, Senem Yıldız Ersoy, Ute Römer-Barron

This study examines the comparative effects of computer-based and paper-based data-driven learning (DDL) on vocabulary acquisition. DDL, which engages learners in exploring authentic corpus data inductively, can be implemented in two ways: through direct interaction with corpora (computer-based DDL) or through teacher-selected concordance materials (paper-based DDL). While computer-based DDL offers learners autonomy, it also presents practical challenges, including learners' unfamiliarity with corpus interfaces and limited technological access in many educational settings. Previous research by Vyatkina (2016) has shown that both approaches can be equally effective for learning German collocations. However, little is known about their effectiveness for English vocabulary learning, which requires the construction of word knowledge from scratch rather than building on existing knowledge (Boulton, 2010). To address this gap, the present study compares the effectiveness and learner behaviors associated with these two DDL approaches in vocabulary teaching.

The participants in the current study are 52 Turkish university students with intermediate-level English proficiency, who were randomly assigned to a computer-based DDL (n = 28) or paper-based DDL (n = 24) condition. Both groups studied 20 target words through inductive DDL tasks over four weeks. Participants in the computer-based DDL group explored concordances directly through the COCA (Corpus of Contemporary American English) or MICASE (Michigan Corpus of Academic Spoken English) interfaces, whereas participants in the paper-based DDL group worked with pre-selected concordance samples (10 per word) chosen by a teacher based on readability, frequency, and usefulness criteria following Reppen (2011). Adopting a quasi-experimental mixed methods design, the study utilized pre-tests and post-tests measuring meaning recall, form recall, and meaning recognition; think-aloud protocols with purposefully sampled participants (n=18); and focus group interviews exploring learner attitudes and behaviors.

Quantitative analyses for the pre-test and post-test results revealed no significant differences between groups on vocabulary gains across all three knowledge levels. These findings suggest that the pedagogical benefits of DDL for vocabulary acquisition can be achieved without requiring direct corpus access, which expands implementation possibilities for educators across diverse technological contexts. However, think-aloud protocols uncovered important qualitative differences in corpus consultation behaviors. Computer-based DDL participants demonstrated greater attention to multiple word knowledge dimensions, particularly frequency, register awareness, and spoken forms through YouGlish integration. They also employed more diverse concordance selection strategies, including choosing texts matching their interests and using extended context features. Paper-based DDL participants completed tasks significantly faster and showed preference for shorter concordances as they believed these provided more concentrated meaning representations.

The study demonstrates that both DDL approaches can achieve comparable learning outcomes while offering distinct pedagogical affordances. The findings provide guidance for corpus integration in language curricula and reveal how learner behaviors and strategies differ across the DDL approaches. For corpus linguistics pedagogy, the results suggest that direct corpus access is not a prerequisite for effective DDL. Overall, this study opens possibilities for broader classroom adoption while maintaining the discovery-learning benefits that characterize data-driven approaches.

#78

Enlivening script to stage: toward a multimodal corpus-based approach to theatre translation studies with Dou E Yuan as an example

Shi Li

Performance- and text-centric views continue to vie for dominance in the theatre translation studies. Despite extensive discussion in theatre translation studies regarding the negotiation of performability and readability, the absence of live performance on the stage substantially complicates any comparison of performative potential across different translations of a given source text. Centering on the tension between performability and readability, this study investigates how seven full English translations of Yuan zaju Dou E Yuan textually envision a stage performance based on a self-built multimodal corpus by integrating Large Language Model-based Text-to-Speech synthesis. The multimodal corpus of English translations of Dou E Yuan built in this study comprises three sub-corpora: text corpus, speech corpus, and visual corpus. The results demonstrate that a multimodal corpus-based approach to theatre translation studies not only helps uncover the interconnection between performability and readability, but also serves to bridge the traditional divide between page- and stage-oriented translation, thereby moving beyond an exclusive focus on linguistic features.

The current study primarily addresses two key issues: first, to investigate the extent to which seven English translations differ on textual, acoustic and visual codes, as well as their impact on theatrical performance; and second, to establish a multimodal approach to theatre translation studies by integrating LLM-based TTS. A translator who focuses solely on linguistic accuracy without considering theatrical viability risks producing a script that sounds correct but feels dead on stage. The findings indicate that by carefully negotiating the complex interplay between performability and readability, translators can significantly enhance the potential for successful staging and reception of translated scripts. Thus, theatrical codes must be considered in tandem rather than in isolation. The multimodal corpus-based analysis highlights the intricate relationship between script and stage presentation, emphasizing the importance of considering verbal and non-verbal features when discussing the theatricality of translated script and its inherent theatrical value.

#79

Changes in referential production among Japanese-English bilingual returnee children: a five-year longitudinal study

Jason Rothman, Maki Kubota, Stefanie Wulff, Vicky Chondrogianni

This five-year longitudinal study examines how dramatic shifts in language exposure shape the referential production of Japanese–English bilingual returnee children after relocation from an English-dominant environment back to Japan. Twenty-five returnees (mean age at return = 9.72 years) were tested at three points—immediately after return, ~1 year later, and ~5 years later—and compared at baseline to age-matched monolingual peers (27 Japanese, 27 English). We focused on how children managed reference in narratives by analyzing their choice of full noun phrases (NPs) and pronouns (overt or null) across three discourse contexts. In First Mention, a new character enters the story, and speakers typically use a full NP (e.g., a boy in English; otokonoko-ga “a boy” in Japanese). In Maintenance, the speaker continues to talk about the same character, where reduced forms are expected—null pronouns in Japanese or overt pronouns in English (e.g., “he ran away” vs. "Ø ran away" in Japanese). In Reintroduction, a character that has been out of focus becomes the topic again, and full NPs are usually re-employed (e.g., “the boy” / otokonoko-wa). These contexts provide a window into how bilingual children balance informativeness and efficiency across their two languages.

Narratives were elicited with wordless picture books (Frog, Where Are You? in Japanese; Frog on His Own in English). Transcripts were segmented into C-units and coded for referential context and form. Exposure was quantified using the BiLEC (Unsworth, 2016), and working memory (WM) was assessed via a N-back task.

At baseline (immediately after return), returnees differed from monolinguals mainly in the Maintenance context: they produced more redundant NPs than both Japanese and English monolingual peers, suggesting a cross-linguistic overspecification strategy. In Reintroduction, returnees used more NPs in English than in Japanese and more than monolinguals, consistent with a clarity-oriented approach in their less dominant language.

Longitudinal results revealed a divergence between the two languages. In English, NP use remained stable over five years across all three contexts, indicating no attrition in referential strategies. In Japanese, however, re-exposure produced measurable changes. In Maintenance, NP use decreased from the second to the third wave, showing greater reliance on context-appropriate null pronouns. In Reintroduction, NP use increased between the first/second and third waves, reflecting more explicit marking when bringing characters back into focus.

Exposure, rather than WM, predicted outcomes in English. One year after return, children with less ongoing English exposure produced more redundant NPs in Maintenance ($r = .48$, $p = .02$). Over the five-year period, greater declines in English exposure likewise predicted increased NP redundancy ($r = .40$, $p = .04$). WM did not account for referential choices once exposure was considered.

In sum, bilingual returnees show (i) stable L2 English referential strategies, (ii) L1 Japanese gains in contexts where reference management is especially demanding—Maintenance and Reintroduction. Overall, the findings underscore the central role of language exposure in shaping bilingual children’s referential strategies, demonstrating that re-immersion fosters more target-like reference tracking in the L1, whereas sustained L2 input mitigates tendencies toward overspecification.

#80

Automatic analysis of second language learners' development of verb-argument constructions in a longitudinal writing corpus

Soyeon Sim

Understanding how second language (L2) learners expand their linguistic repertoire over time has been a central question in second language acquisition (SLA) research. A substantial body of learner corpus research has examined L2 use and development through cross-sectional corpora spanning different proficiency levels. However, such group-based designs assume homogeneity within levels and overlook individual variability as well as the dynamic nature of L2 development over time (Lowie & Verspoor, 2019). To address these limitations, dynamic usage-based studies have often adopted longitudinal designs that trace individual learner trajectories, typically through small-scale case studies (Verspoor et al., 2020). Grounded in the dynamic usage-based approach and drawing on a large-scale longitudinal corpus, the present study investigates the longitudinal development of selected verb-argument constructions (VACs) in L2 writing and tracks their individual trajectories over time. The study is guided by the following research questions:

1. To what extent do L2 learners of English expand their repertoire of VACs over time in a longitudinal L2 learner corpus?
2. How does the distribution of verbs in VACs evolve across different time points in L2 learners' writing?
3. How do L2 learners' VAC usage patterns differ from those of successful upper-level academic students in an American university?

The learner data came from the Longitudinal Database of Learner English (LONGDALE), consisting of academic writing by learners of English, and reference data were drawn from the Michigan Corpus of Upper-level Student Papers (MICUSP). Ten VACs were selected from Goldberg (1995) as target VACs, including the passive, ditransitive, and caused-motion constructions. These constructions were automatically identified and extracted using Python's NLP library spaCy, leveraging dependency relations and syntactic features. The resulting VAC frequencies and verb types formed as the basis for subsequent statistical analyses.

A linear mixed-effects model was employed to examine VAC frequency and diversity changes over time in the LONGDALE data. Collostructional analyses were also conducted to examine which verbs were attracted to the target VACs and whether there were any changes over time. Furthermore, the VAC frequency and diversity of L2 writing were compared to the writing of L1 English speakers. Findings revealed non-linear changes in frequency for most target VACs, with only the passive construction showing a significant increase over time. The overall VAC diversity, on the other hand, suggested a significantly increasing trend across the three years. Collostructional analyses revealed that developmental changes in VAC usage, with near-Zipfian distributions of items in the verb slot and an increase in verb type diversity at later stages. A comparison with L1 academic writing (MICUSP data) indicated that L1 writers exhibited both higher frequencies and greater diversity in VAC usage compared to L2 writers. Overall, the results highlight the dynamic nature of L2 learners' longitudinal VAC development in academic writing, with divergent patterns across target VACs but a consistent increase in the diversity for VAC usage.

#81

Validating the Phraseological Complexity Analyzer (PaCa): a digital tool for assessing phraseological diversity and sophistication

Shuyuan Tu, Daniel H. Dixon

Phraseological complexity has been proposed as a critical construct in L2 development (Bestgen & Granger, 2014; Biber et al., 2011; Paquot, 2019). Research suggests that phrasal development follows a gradual but uneven trajectory that varies across proficiency levels and learner groups (Bestgen & Granger, 2014; Siyanova-Chanturia, 2015). Given this variation, L2 instructors and researchers face challenges in assessing their students' phraseological L2 development. These challenges suggest a need for an automated approach that can systematically measure the level of phraseological complexity in L2 learners' output. Towards addressing this need, the proposed Phraseological Complexity Analyzer (PaCa) is an NLP-powered research tool designed to automate the analysis of the development of phraseological complexity in L2 learners' written and spoken output (as machine-readable texts) following Paquot's (2019) operationalization of phraseological complexity. That is, PaCa approaches phraseological complexity by targeting three phraseological constructions: adjectival modifiers, adverbial modifiers, and direct objects, and the tool programmatically extracts and quantifies the pairs identified. PaCa provides various preprocessing options for the input texts, allowing users to select lemmatization, parts-of-speech (POS) tagging, and dependency parsing before the analysis. After preprocessing, PaCa extracts the word pairs with their raw and normalized frequencies (per 1,000 words) and saves the results as downloadable lists.

Two dimensions of phraseological complexity are measured: phraseological diversity and phraseological sophistication. Phraseological diversity, which reflects the range of phraseological structures learners use, is operationalized as the ratio of unique phraseological units to the total number of phraseological units using the root-type-token ratio (RTTR) for word pairs within the three target constructions. Phraseological sophistication, which captures the degree of sophistication of these phraseological units, is measured using two approaches. The first approach is comparing with the Academic Collocation List (Ackermann & Chen, 2013), and the second approach is through average Pointwise Mutual Information (PMI) scores, which measure the strength of association between words in the word pairs against their expected strength of association in a comparable reference corpus.

The accuracy of the tool was investigated by processing a subsection of the EF-Cambridge Open Language Database (EFCAMDAT; Geertzen et al., 2013), an open-access corpus containing written data produced by English L2 learners from a range of L1 backgrounds. Results from a 100,000-word sample of L1 Mandarin learner data from intermediate to advanced levels (B1-C1) were used to demonstrate the tool's use and functionality. The analysis revealed a consistent increase in phraseological diversity across all three constructions from B1 to C1. However, in terms of phraseological sophistication, only dobj showed a statistically significant increase from B2 to C1, while no differences were observed for amod or advmod. The accuracy was reported using precision (≥ .87), recall (≥ .87), and F1 scores (≥.92), based on manual assessments of the target phraseological constructions in a random sample of texts. Assessments suggest that PaCa has a high level of accuracy in both tagging and extracting the target constructions. Moreover, its user-friendly design makes it accessible to a broad audience in tracking learners' phraseological development.

#87

Thematic progression anomalies in Japanese university students' argumentative essays: a functional linguistic perspective

Tsukuru Kamiyama

Organizing information in academic writing is one of the most important skills for beginning academic writers to learn, and acquiring it is highly demanding. One factor making it challenging is how information is organized in academic writing—it is distinct from how it is managed in our daily communications. This hurdle is particularly daunting for non-native English writers to overcome (Hyland, 2004; Schleppegrell, 2004).

One of the ways to manage discursive flow in academic writing is by strategically choosing what to place at the beginning of the sentence, a linguistic operation called thematization, and this strategy has been extensively scrutinized by using the concepts called Theme and Rheme (Halliday & Matthiessen, 2014). Prior research (e.g., Daneš, 1974) has identified three general Theme-Rheme relations, or thematic progression patterns, and their respective function in terms of information flow, finding that experienced writers utilize the three patterns depending on their purpose while language learners use them ineffectively and often result in deviating from the patterns. However, deviation from the three Theme-Rheme patterns is not unique to novice writers; in fact, it is often observed in experts' writing as well. For instance, McCabe (1999) analyzed published history texts in terms of thematic progression, reporting incidences that did not fall into the three patterns and classifying the anomalies into four groups (pragmatic, grammatical, extraposed, and metatextual) based on their characteristics. Similarly, Hawes (2010) focused on the irregularity with thematic progression in published British news articles, identifying five categories of abruptive thematic progression, or "non-participant breaks." These categories are WH- & polar interrogatives, "It" and "There" predicates, verbal group breaks, bound clause breaks, and annexes, and Hawes concluded that they were used to (i) change the course of discourse and (ii) serve the author's evaluative purposes. Despite their usefulness and contribution to better understanding of thematic progression and to demystifying academic writing, such studies are scarce: more research needs to be done in other disciplines than communication and history and different genres than news articles and expository texts with novice / lower-proficiency learner populations.

To this end, this study investigates 111 argumentative essays written by 67 Japanese university students at intermediate levels and 44 native English language teachers in terms of their deviations from the three thematic progression patterns, using a Corpus-Assisted Discourse Analysis approach. The data come from the International Corpus Network of Asian Learners of English (ICNALE) (Ishikawa, 2023), a publicly available corpus. The corpus was selected for its unique characteristics: the language learner group is Japanese nationals, who have not been studied on this issue; and both the L1 and L2 writers in the corpus wrote essays under exactly the same conditions (i.e., timed, on the same topic, and with the same word limit), which makes possible a more direct comparison between the two groups. This ongoing study will discuss the results and shed light on when and how to deviate from the three thematic patterns while maintaining a comprehensive discourse flow, thereby offering implications for better teaching L2 writers how to organize discourse flow.

#88

Lexical organization in learner production: the role of frequency, context, and semantics

Anton Vegel

Word frequency (WF) has had a strong influence on how we understand and analyze natural language. However, as a repetition-based metric, its reliance on aggregate counts without consideration for context is a considerable weakness. Unsurprisingly, however, WF has been shown to be a robust predictor of word processing and production of both words (Hamrick & Rebuschat, 2014) and constructions (Ellis & Ferreira–Junior, 2009). In contrast, context-based metrics look beyond aggregate counts and consider the role of words in contexts. These metrics align with the rational analysis of memory (Schooler & Anderson, 1997) and thus potentially offer a more plausible account of lexical organization. Context-based metrics capture the number of contexts a word appears in (contextual diversity; Adelman et al., 2006), the co-occurrence probability of those contexts (contextual distinctiveness; McDonald & Shillcock, 2001), or the variability of semantic content in those contexts (semantic diversity; Hoffman et al., 2013; semantic distinctiveness; Johns et al., 2016; Jones et al., 2012). Furthermore, contextual diversity and distinctiveness, as well as semantic diversity and distinctiveness, have been shown to explain lexical processing beyond WF (Adelman & Brown, 2008; Hamrick & Pandža, 2020; Jones et al., 2012). However, findings have not been wholly consistent. While research investigating such metrics have utilized natural language via learner production data in the past (Berger et al., 2019; Crossley et al., 2013)many have relied on behavioral data such as psycholinguistic norms and reaction time in their own right (Chang et al., 2023; Hashimoto & Egbert, 2019).

This study thus aims to utilize recent developments in lexical metrics and extend the research lens back to natural language data. For this study, the large-scale learner corpus, EFCAMDAT (Shatz, 2020), served as a data source subsampled to balance task, proficiency, and first language groups. Learner data was used to create ten lexical metric representations: five WF and five context-based metrics each. These were calculated following standard procedures for word indices (see Kyle & Crossley, 2015). These metrics were selected based on past studies and recent developments (i.e., recent social media-based WF norms; Herdağdelen & Marelli, 2017; and context-based metrics built with a usage-based assumption from social media data using the individual as an organizing principle; Johns et al., 2020; Johns, 2021). To allow for a balanced comparison of the variance captured by the metrics, only content words shared across all lexical metric dictionaries were used for analysis with strict control for word count outliers. Preliminary multiple linear regression analyses, using R-squared and semi-partial delta R-squared, show that context-based metrics explain more variance in learner production data than WF metrics across both broad and fine-grain proficiency level variables. Furthermore, the unique variance explained by WF metrics diminishes when context-based metrics are included in the model. These findings align with recent confirmations of the value of context-based metrics (Kumar, 2020; Norman, 2025) as well as a general call to embrace measures beyond WF (Gries, 2024).

#91

Pronoun use in compliments: capturing cultural variation in human experience

Jini Jung

This study examines how pronoun use in compliments—universal yet culturally variable speech acts—captures variation in human experience through language. Compliments, which inherently require both evaluation by the speaker and a response from the addressee, reveal how linguistic choices reflect cultural values and communicative expectations (Abramova, 2015). Pronouns, in particular, highlight perspective-taking and social alignment, so they can make the cultural variation of interaction visible.

This research investigates how native speakers of English and Korean use subject pronouns ("I," "you," "it") in compliment settings, focusing on frequency and sequential placement. Data were collected through discourse completion tasks (DCTs): 66 participants (30 L1 English speakers and 36 L1 Korean speakers) each responded to 10 everyday compliment scenarios, yielding 660 elicited compliments with 1003 tokens. This small but systematic corpus illustrates how DCT-generated datasets can be analyzed with corpus-informed methods.

The data analysis demonstrated cross-linguistic differences. English speakers strongly preferred first-person self-reference ("I like your jacket"), while Korean speakers actively avoided it and employed alternative structures ("Your jacket is pretty"). Sequentially, English speakers often opened compliments with "I" while Korean speakers either delayed or omitted self-reference. A non-parametric Mann–Whitney U test confirmed that the difference between English and Korean groups was statistically significant ($p = 5.561e-05$) with a large effect size (Cohen's $d = 1.33$, 95% CI [0.71, 1.95]).

The findings indicate that the observed differences in pronoun use during compliments reflect divergent pragmatic norms, which stem from broader cultural orientations of individualism in English and collectivism in Korean. This aligns with Kashima and Kashima's (1998) argument that cross-linguistic patterns in pronoun use reflect underlying cultural norms. These results show that even subtle grammatical choices can reveal cultural distinctions in interpersonal experience, underscoring the value of a language-science approach to everyday speech acts.

#92

A computational diachronic analysis of Gen-Z Mental Health Discourse: a large-scale reddit corpus study from pre- to post- COVID

Felix Mao

Throughout history, all generations have faced mental health challenges shaped by their sociohistorical context. Generation Z, however, faces distinct challenges influenced by digital saturation and the global disruption of the COVID-19 pandemic, leading them to turn increasingly to online platforms to express distress and seek support. These circumstances have shaped how mental health is discussed among Gen Z, producing distinctive lexical choices and discourse strategies for sharing experiences. While discourse analyses of mental health have largely relied on qualitative approaches and focused on the general population, large-scale, corpus-driven computational investigations specific to Gen Z's linguistic behavior in social media remain underexplored. To address this gap, we construct and analyze what we believe to be one of the first and largest dedicated corpora for studying Gen Z mental health discourse online. The dataset is drawn from 11 subreddits using python webscraping, including mental health support spaces and Gen Z identity forums. Crucially, "Gen Z" was defined by behavioral cross-posting: users active in both mental health and Gen Z-identified communities were classified as Gen Z, while those active only in mental health forums served as a non-Gen Z control group. This approach yielded a corpus of more than 3 million posts and comments from over 320,000 unique users spanning 2017–2025. Given the scale and complexity of the data, we employ a hybrid methodology integrating statistical corpus linguistics with machine learning-based algorithms to ensure both interpretability and analytical depth. A diachronic keyness analysis across pre-, during-, and post-COVID corpora was conducted to determine lexical features that characterized mental health conversations across subreddits and time periods. In addition, using the Python NLTK (natural language toolkit) and scikit-learn packages, sentiment analysis and topic modeling were applied to uncover underlying thematic structures, trace clusters of lexical items, and examine syntactic and semantic patterns in Gen Z discourse over time. This design enables a robust analysis of mental health discourse both diachronically and through comparison with non-Gen Z users. Our findings reveal three key insights. First, within the Gen Z cohort, we identify patterns of ritualized support and the use of highly negative disclosures as signals that drive in-group engagement—a phenomenon we term negative in-group authentication. Second, temporally, the analysis shows that the pandemic reshaped the framing of pre-existing topics rather than introducing new ones, followed by a post-pandemic downturn in sentiment that suggests a delayed cumulative impact. Third, a comparative analysis demonstrates fundamental differences between groups: Gen Z discourse tends toward abstract and existential concerns, while non-Gen Z discourse is more grounded in severe physical symptoms. By introducing a novel corpus and replicable computational framework, this study advances both methodological and substantive understanding of youth mental health discourse online. Beyond its analytic contributions, our findings highlight how global trauma reconfigures linguistic expression among digitally native youth and underscore the need for digital mental health interventions, platform design, and AI tools that are culturally attuned, linguistically informed, and capable of supporting rather than pathologizing Gen Z's modes of expressing distress.

#93

Do keywords tell a different story?
A comparison of key feature and keyword analyses in research article introductions

Nergis Danis

This study compares two corpus-based keyness methods, key feature analysis (Egbert & Biber, 2023) and text dispersion keyword analysis (Egbert & Biber, 2019), to investigate how disciplinary variation is realized in research article Introductions. Using the Quantitative Research Articles Corpus (Q-RAC), which contains 900 empirical articles from six disciplines, the study focuses on two disciplines: Applied Linguistics and Mechanical Engineering. For each discipline, its Introduction subcorpus was contrasted with reference corpora comprising the Introductions of the remaining disciplines to identify distinctive grammatical features (i.e., key features) and salient lexical items (i.e., keywords). The comparative functional analysis of the resulting key features and keywords demonstrates that the two methods yield complementary insights. Key features highlight broader discourse and rhetorical functions, some of which are shared across disciplines but realized through different forms. Keywords, on the other hand, foreground domain-specific lexis dispersed across part-genres. Applied Linguistics Introductions emphasize terms tied to language learning and teaching, while Mechanical Engineering Introductions highlight terminology related to materials, mathematical concepts, and spatial descriptions. Although a small subset of keywords also signals rhetorical functions (e.g., reviewing previous literature in Applied Linguistics; identifying a problem/gap in Mechanical Engineering), the primary contribution of keyword analysis lies in revealing discipline-specific terminology. Taken together, these findings show that key feature and keyword analyses answer different but complementary questions: the former reveals functional tendencies across disciplines, while the latter highlights the lexical items that index disciplinary signatures. Using both approaches in tandem provides a more comprehensive account of disciplinary writing patterns, with implications for corpus-based genre analysis and academic writing instruction.

#95

Linguistic differences in humor: a feature-based comparison between human and AI-generated jokes

Freya Pan

Background. Humor is one of the most creative and socially embedded forms of human communication, reflecting both emotional expression and cultural interaction. With the rapid development of generative artificial intelligence, the differences between human and AI humor have become an important issue in computational humor research. However, existing comparative studies on human and AI humor face two major limitations: first, most rely on small-scale corpora or subjective ratings, making it difficult to uncover systematic linguistic differences; second, even when linguistic features are considered, prior analyses are often impressionistic or confined to a single dimension, failing to explain the deeper mechanisms underlying human-AI humor differences. Without systematic research, we cannot determine how far AI still is from human humor, which may result in continued constraints on its ability to replicate and simulate human humor.

Aims. This study aims to adopt a quantitative linguistics approach, using a data-driven paradigm to systematically compare human- and AI-generated humor across multiple linguistic dimensions, including lexical features, syntactic structures, sentiment, semantic patterns, and prosodic features.

Samples. A large-scale parallel humor corpus was first constructed, consisting of 22,000 human-authored jokes (sourced from Reddit, short jokes, and puns) and 66,000 AI-generated jokes (produced by GPT-4.1-mini, Llama-3.3-70B, and DeepSeek-V3-Chat).

Methods. An automated program was developed to extract and quantify linguistic indices, and Elastic Net logistic regression was employed to identify the most discriminative linguistic features between human and AI humor. This approach overcomes the shortcomings of earlier studies with respect to corpus size and insufficient data-driven methods.

Results. The results show that AI-generated jokes rely more heavily on connectives and retrospective narration. By contrast, human-authored jokes exhibit greater lexical density, conditional subordination, emotional diversity, semantic incongruity, and thematic consistency, as well as more frequent use of nouns, interrogatives, social references, affective markers, and deliberate rhythmic devices. These differences highlight the significant gap between human and AI humor in terms of creativity, emotional depth, and prosodic design.

Conclusions. These findings reveal important linguistic differences between human and AI-generated humor, offering empirical insights that may support the theoretical study of humor mechanisms. The study hopes to contribute both to the theoretical validation of humor mechanisms and to the practical advancement of explainable and human-like humor generation in LLMs.

#96

Mapping narratives of resilience in the Heat–Health–Climate nexus: insights from a corpus linguistics approach

Ersilia Incelli

This study explores the intersection of heat, health, and climate change through a corpus linguistics approach, examining how these issues are framed and discussed in two distinct corpora: one derived from media outlets, mainly English language newspapers, and the other from international scientific journals regarding the environment and health, for example Sustainable Environment. By applying a corpus-assisted methodological procedure, involving keyword analysis and collocation patterns, reflecting strategic linguistic choices such as transitivity, and nominalization, the research aims to identify dominant themes, terms, and discourse structures associated with heat-related health risks in the context of climate change (Fløttum, et al., 2016). The analysis will have a particular focus on urban contexts and communities, identifying frequent and significant collocations, such as heat stress, heat stroke, heat-related mortality, urban heat, extreme weather events, public health, the elderly, and their association with other clusters like building resilience, climate/urban adaptation, heat action plans, marginalized/vulnerable populations, policy response. Additionally, the study will compare how the language of urgency, responsibility, and mitigation is employed in both corpora. Findings from this type of analysis can provide insights into how the framing of heat-health-climate issues differs across media and scientific discourse; for example, in media coverage, how much is heat-related health risk framed as an urgent public issue, or is it a scientific concern for policy makers? Thus, the research attempts to contribute to an enhanced understanding of the role language plays in shaping public and expert perceptions of climate-related health challenges (Bortoluzzi and Zurru, 2024).

References

Bortoluzzi M. & E. Zurru eds. (2024). Ecological Communication and Ecoliteracy: Discourses of Awareness and Action for the Lifescape. London: Bloomsbury Academic.

Fløttum, K., Dahl, T., & Rivenes, V. (2016). Narratives in Climate Change Discourse. Wiley.

#98

Context valence as a tool for categorising semantic prosody: investigating register-sensitivity and the impact of polysemy

Mathias Russnes

This paper investigates how effectively different measures of context valence can be used in studies of semantic prosody. Semantic prosody is a well-known concept within corpus linguistics, and describes how neutral items recur in evaluative contexts (Sinclair 1996; Stewart 2010), a tendency which has been found to be sensitive to both polysemy and register (Bublitz 1996; Tribble 2000). However, research on semantic prosody has been criticised for its reliance on manual analysis, due to potential subjective influence and issues of replicability (Dilts & Newman 2006; Bednarek 2008; Winter 2019). To address this, the more objective measure context valence was introduced (Snefjella & Kuperman 2016), along with the modified metric absolute context valence (Winter 2016, 2019). The aim of this paper is to test the effectiveness of these measures – together with the new metric weighted context valence – in the categorisation of semantic prosody, as well as their application in examining the influence of register and polysemy on the concept. The results, based on a study of the lemma commit in four registers of BNC2014, imply that the measures can effectively identify the impact of polysemy, as well as register-variation. Although the metrics exhibited varying degrees of precision in categorising individual instances, they were consistently more reliable in identifying negative than positive contexts.

References

Bednarek, M. (2008). Semantic preference and semantic prosody re-examined. Corpus Linguistics and Linguistic Theory, 4(2), 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.

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, 135-146. https://www.sciencedirect.com/science/article/pii/S0010027716301792

Stewrt, D. 2010. Semantic Prosody. A Critical Evaluation. London: Routledge.

Tribble, C. 2000. 'Genres, Keywords, Teaching: Towards a Pedagogic Account of the Language of Project Proposals'. In L. Burnard and T. McEnery (eds.) Rethinking Language Pedagogy from a Corpus Perspective, pp. 75-90. Frankfurt am Main: Peter Lang.

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.

#99

The sociolinguistic impact of divergence in human experience: variation and change in Laurentian French

Hélène Blondeau, Raymond Mougeon, Mireille Tremblay

INTRODUCTION. This corpus-based study of two varieties of Canadian French combines the tools of micro-diachronic sociolinguistics and urban dialectology to better understand how divergence in human experience leads to language change. We present results from a project which studies two genetically related varieties of Canadian French: one spoken in Montréal, Québec, a metropolis where Francophones are in the majority, and the other spoken in Welland, Ontario, a much smaller city where Francophones represent only 10% of the population. In this paper, we contrast the linguistic and social patterning of three sociolinguistic variables across the two dialects over 40 years.

METHODS. Patterns of linguistic change are documented via a comparison of corpora collected in each community in the 1970s and the 2010s. All four corpora are stratified with respect to gender, age and socio-economic status (SES) (and bilingualism in Welland). This unique dataset allows for the comparison of the two sister varieties at two points in time, in the 1970s and the 2010s.

RESULTS Our apparent- and real-time analyses show three distinct patterns of variation, all demonstrating increased divergence between the two varieties in the context of standardization. The first variable “automobile” shows that, overall, in both communities standard variants auto and voiture have increased at the expense of vernacular variants char and machine. However, the speech of the younger generations reveals a pattern of divergence, namely, sharp decline of char and sharp rise of auto in Welland and resilience of char in Montreal. A trend toward increased standardization is also observed in the case of the second variable “pronouns”, with standard simple forms (nous, vous, elles, eux) replacing the traditional vernacular complex variants (nous-autres, vous-autres, eux-autres) in both communities but in this case more so in Montréal than in Welland. Finally, in case of the third variable “consequence markers”, the standard forms (alors and donc) compete with the vernacular ça fait que, traditionally reduced to [fak]. Both communities show a loss of the vernacular variant [fak]. In Welland, this traditional variant is replaced by the English borrowing so, while in Montréal, the change is more subtle with [fak] being replaced by a less stigmatized pronunciation [fɛk]. In addition, we observe a sharp decrease of alors in favor of donc in Montréal, but the opposite effect in Welland, where alors is more resilient. In sum, while both communities show signs of standardization over time, there is a divergence in the inventory of variants and the pace of change.

CONCLUSION. We interpret these linguistic differences in light of community-specific social changes: 1- the Quiet Revolution in Quebec, increased schooling, and contact with other French dialects in Montreal; and 2- contact with English, decreased intergenerational transmission, and the increased role of schools in the transmission of French in Welland. Because it examines the relationship between community trends and socio-historical and demolinguistic contexts, our study show how comparative micro-diachrony emerges as a useful tool to tackle one of the most important challenges in sociolinguistics : the Actuation Problem (Weinreich et al., 1968).

#108

Leveraging masked language models to measure association strength in contiguous and dependency bigrams

Hakyung Sung, Kristopher Kyle

In language production research, phraseological competence (i.e., the ability to use naturally co-occurring word combinations) has been recognized as a core dimension of language proficiency (Nation, 2001; Römer, 2009). In second language (L2) studies, this competence has often been assessed through n-grams, with findings showing that more proficient L2 users tend to produce n-grams that are strongly associated in the target language, typically measured using metrics such as pointwise mutual information (MI) (Granger & Bestgen, 2014). More recently, a grammar-informed approach has been proposed, adding a layer of lexicogrammatical relations (e.g., verb + direct object), which may include non-adjacent elements (Paquot, 2019; Kyle & Eguchi, 2023).

To extract these indices, prior studies relied on large reference corpora to extract phraseological indices. For example, Paquot (2019) computed dependency-based MI scores from the L2 Research Corpus (L2RC), matched VESPA learner pairs, and then calculated per-text mean MI values. Kyle and Eguchi (2023) similarly used the spoken section of COCA (Davies, 2009) to estimate association strength for contiguous and dependency n-grams. Although these approaches have shown predictive success for L2 proficiency, methodological limitations remain: Results might be sensitive to the choice of reference corpus (Paquot, 2019, p. 137) and n-grams are often evaluated outside their larger co-text and context.

To address these limitations, this study proposes an estimate of phraseological association using a masked language model. For each sentence, we score adjacent bigrams with a left-to-right pseudo log-likelihood: (i) mask the first word and the remainder to get log P(w_i | left context); (ii) mask the second word and the remainder to get log P(w_{i+1} | left context + w_i). Their sum is the bigram's in-context score. For example, in "I drank strong tea this morning," log P(strong | I drank) + log P(tea | I drank strong) should presumably exceed the score for "powerful tea," reflecting bigram preferences in the language models. We also annotate POS and direct dependencies (e.g., verb-dobj) to analyze both contiguous pairs and grammar-restricted subsets. The method preserves sentence context and can be domain-adaptively fine-tuned (e.g., conversational corpora) to align scores with the target register.

In the experiment, we extracted contiguous and dependency bigrams and computed in-context association scores using three masked language models: BERT-uncased (Devlin, 2018), a conversational BERT (https://huggingface.co/DeepPavlov/bert-base-cased-conversational) , and a BERT-uncased fine-tuned on MICASE (Römer, 2017). We then compared these model-based indices with legacy corpus-based indices (Kyle & Eguchi, 2023; computed against COCA-Spoken) on the NICT-JLE corpus (Izumi et al., 2004) by correlating each index with L2-English oral proficiency. Preliminary results are promising for L2 oral proficiency prediction. For contiguous bigrams, the masked-LM-based average log-probability correlated more strongly with proficiency than a leading corpus-based comparator (MI2): r = .643 vs. .545. Among the four dependency relations examined (i.e., verb-direct object, verb-subject, verb-advmod, noun-amod), verb-advmod and noun-amod outperformed legacy measures; e.g., in amod, the model-based index surpassed ΔP_depcue (r = .459 vs .116). These zero-order gains suggest added value in multivariate models alongside other lexicogrammatical predictors.

#114

What are hybrids?
A multidimensional analysis of hybrid texts on the French and Swedish web

Saara Hellström, Erik Henriksson, Veronika Laippala

Linguistic variation on the web is immense and extensive. Biber & Egbert (2018) have shown that the open English web abounds not only with recognizable registers, i.e., situationally bound, culturally recognized language use with a purpose (Biber & Egbert 2023), but also with texts that combine features of different registers, i.e., hybrids (Biber & Egbert 2018). Previous research has shown that hybrids are salient as they make up a notable part of web language use (Biber & Egbert 2018; Repo et al. 2021). Nevertheless, hybrids have not been extensively explored, most likely in lack of data that would include (enough) texts representing different register combinations. Moreover, although the internet is a multilingual space, extensive research in web language use in other languages than English remains scarce.

In this paper, we aim to fill these gaps by exploring hybrid texts on the French and Swedish web in HPLT 2.0 data (Burchell et al. 2025) collected from the unrestricted web in a data-driven manner without any preselected criteria. We aim to answer the following questions: 1) what are the linguistic characteristics of hybrids?, 2) how do hybrids relate to non-hybrids?, 3) what are the cross-linguistic (dis)similarities between French hybrids and Swedish hybrids?

Our data consists of two 30,000-document samples, one from each language, divided into 30 categories of which 9 are French and 12 Swedish hybrids. We apply multidimensional analysis (MDA; Biber 1988) in which correlating linguistic features form dimensions of linguistic variation that can be used to characterize and compare texts (i.e., hybrids and registers). As features, we include all pos tags, morphological features and dependency relations provided by the dependency syntax parser Trankit (Nguyen et al. 2021) following the UD scheme. Then we apply feature selection for the final set of features. First findings indicate that hybrids show some language-specific features and that they deviate from the registers of which they combine features.

References

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

Biber, D. & Egbert, J. (2023). What is a register? Accounting for linguistic and situational variation within – and outside of – textual varieties. Register Studies, 5(1), 1-22. https://doi.org/10.1075/rs.00004.bib

Burchell, L. et al (2025). An expanded massive multilingual dataset for high-performance language technologies (HPLT). arXiv. https://doi.org/10.48550/arXiv.2503.10267

Nguyen, M. V. et al. 2021. Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. In Dimitra Gkatzia & Djamé Seddah (eds.), Proceedings of the 16th conference of the European chapter of the Association for Computational Linguistics: System demonstrations, 80–90. Association for Computational Linguistics. https://arxiv.org/pdf/2101.03289

Repo, L. et al. 2021. Beyond the English web: Zero-shot cross-lingual and lightweight monolingual classification of registers. In Ionut-Teodor Sorodoc, Madhumita Sushil, Ece Takmaz & Eenko Agirre (eds.), Proceedings of the 16th conference of the European chapter of the Association for Computational Linguistics: Student research workshop, 183–191. Association for Computational Linguistics. https://arxiv.org/pdf/2102.07396.pdf

#120

Linguistic creativity and spontaneity: an investigation of hypothetical *if*-clauses

Yen-Po Chen, Siaw-Fong Chung

Linguistic creativity is generally understood as the capacity to use language in novel and effective ways (Carter, 2016; Cook, 2000). It often emerges through the inventive recombination of established words, phrases, or constructions to produce witty or striking effects. This research explores how such creativity is manifested in the r/Showerthoughts subreddit, a register characterized by spontaneity, where language use is self-contained, humorous, and sometimes bizarre. Drawing on a compiled corpus of r/Showerthoughts (Authors, 2019; To appear), this study focuses on hypothetical if-constructions, specifically "If…were", which allow imaginative scenarios and creative reuse of familiar elements. The target constructions were extracted using a corpus analysis tool, and sentences were sorted according to shared features. There are around 2,500 instances of this type of if-clause in the corpus. These clauses are notable because they frequently recycle commonly recognized elements; for example, there are 12 instances of "If Adam and Eve were…". They combine such elements in unexpected or humorous ways by creating playful links between the protasis and the apodosis of if-clauses. The analysis contributes to a broader understanding of how linguistic creativity operates in online contexts, highlights how hypothetical if-clauses are used to construct imaginative scenarios, and shows how shared aspects of human experience shape their spontaneous expression.

#121

Clustering embeddings from register classifiers reveals fine-grained structure within web registers

Erik Henriksson, Tuomas Lundberg, Antti Kanner, Veronika Laippala

Registers – language varieties marked by co-occurring linguistic patterns in particular contexts (Biber & Conrad 2019) – are typically described in broad categories, but texts within the same register often differ in systematic ways. Such differences can reflect functionally distinct subregisters, topical groupings, or gradience along register continua (Biber & Egbert 2023), yet identifying these patterns often requires manual, corpus-specific analysis. This paper introduces a computational method for discovering fine-grained structure within web register categories by combining supervised register classification with unsupervised clustering. We explore what kinds of patterns emerge from this approach, and whether such patterns hold across languages. We fine-tune multilingual XLM-R models (Conneau et al. 2019) on a 25-class register taxonomy (Henriksson et al. 2024) to classify one million documents per language from English, Finnish, and Swedish web corpora sampled from the HPLT dataset (Burchell et al. 2024). This yields large register-specific datasets (e.g. approximately 490,000 Narrative Blogs, 79,000 Interactive Discussion pages), with 5-15% of documents assigned to multiple registers depending on language. Clustering the fine-tuned embeddings reveals both topical and functional subtypes. Topically, Finnish documents combining Narrative Blog + Opinion separate into consumption-focused versus culture-focused subtypes, while sports-focused clusters emerge from English Interactive Discussion documents. Functionally, clustering captures, for instance, gradient variation in interactivity: across all three languages, Narrative Blog texts separate by degree of reader interaction, with blogs containing comment sections forming distinct clusters that blend narrative and interactive features. These findings align with recent work on the continuous nature of register variation (Biber & Egbert 2023) and provide a data-driven method for identifying where texts shift along register continua. We validate our approach through UMAP visualization (McInnes 2018) and keyness analyses, finding that visually separated clusters have distinctive linguistic features. SVM classifiers trained to predict cluster membership achieve 95-99% test accuracy, demonstrating that subtypes represent learnable patterns. Crucially, clustering baseline (non-finetuned) XLM-R embeddings produces less linguistically distinctive clusters that group texts more by document formatting and boilerplate language than functional patterns. Fine-tuning on register classification thus reorganizes embedding space to capture topical and functional variation at finer granularity than the training taxonomy.

References

Biber, D., Conrad, S. 2019. Register, Genre, and Style. Cambridge: Cambridge University Press.

Biber, D., Egbert, J. 2023. “What is a register? Accounting for linguistic and situational variation within – and outside of – textual varieties”. Register Studies 5:1, 1-22. https://doi.org/10.1075/rs.00004.bib

Burchell, L. et al. 2025. “An expanded massive multilingual dataset for high-performance language technologies (HPLT)”. https://doi.org/10.48550/arXiv.2503.10267.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V. 2019. “Unsupervised cross-lingual representation learning at scale”. https://doi.org/10.48550/arXiv.1911.02116.

Henriksson, E., Myntti, A., Hellström, S., Eskelinen, A., Erten-Johansson, S., Laippala, V. 2024. “Automatic register identification for the open web using multilingual deep learning.” https://doi.org/10.48550/arXiv.2406.19892.

McInnes, L., Healy, J., Melville, J. 2018. “Umap: Uniform manifold approximation and projection for dimension reduction”. https://doi.org/10.48550/arXiv.1802.03426.

#123

Statutory interpretation of a verb using prototype-by-component analysis

Alyssa Lowry

In Pierre-Noel Ex Rel. KN v. Bridges Public Charter School in 2023, an important linguistic question was posed: does the ordinary meaning of transportation in the Individuals with Disabilities Education Act (IDEA) extend to a pedestrian mode of transport? The District Court of Columbia used the sole search term transportation in the Corpus of Historical American English (COHA; Davies, 2010) from 1965-1975, finding no concordance lines referring to a pedestrian mode of transportation. The Court concluded that transportation does not include pedestrian means. However, the Court only examined ordinary meaning around the time that IDEA was established, yet Lee & Mouritsen (2017) argue that ordinary "public meaning" merits examining contemporary evidence when "contemporary reliance interests and fair notice" are in question (p. 825). In contrast, the current study includes the use of the News on the Web (NOW) corpus (Davies, 2016) to examine contemporary data with transport* as a search term to yield all results for the lemma TRANSPORT. Initial corpus queries yielded limited evidence for a pedestrian means of transportation, such as this U.S. concordance example: "primary mode of transport will be by foot" (qtd. in Davies, 2016).

Furthermore, Gales & Solan (2019) assert that merely looking for attested referents of a term in a corpus excludes potential referents that only exist outside the corpus. This is known as the "blue pitta problem" because the absence of blue pitta (a bird of Asia) in COCA does not negate its true membership to the category bird (p. 500). One potential solution is using PCA (Egbert & Lee, 2024), which combines aspects from prototype theory and componential analysis to compare features among lexical categories and (potential) category members. However, the examples in Egbert & Lee (2024) only include concrete nouns. The current study demonstrates how PCA can be applied to a verb—transport. To address the added complexities involved in performing feature analyses for verbs, principles from both componential analysis (Katz & Postal, 19964) and frame semantics are explored (Fillmore, 1994; Boas, 2008).

Following the framework of Egbert & Lee's PCA (2024) analysis, the operative words include the lemma TRANSPORT for the category and by|on foot as a potential category member. For comparative analysis, synonyms for TRANSPORT were found using thesauruses and corpus searches; category members for TRANSPORT and synonyms were found using corpus methods. The NOW corpus (Davies, 2016) was selected as an appropriate corpus for interpreting ordinary meaning of English terms around the time of litigation in Pierre-Noel Ex Rel. KN v. Bridges Public Charter School. Collocate analyses analogous to those in Egbert & Lee (2024) were performed to identify semantic components; principles from frame semantics were also explored to augment componential analyses. It is hypothesized that a pedestrian means of transportation will be a member of the category TRANSPORT, albeit not highly prototypical. This research demonstrates that PCA is a helpful approach for statutory interpretation and can be applied to additional parts of speech.

#124

The role of register range in explaining lexical decision times

Daniel Keller, Earl Brown, Brett Hashimoto

Research has shown that words encountered across texts and in diverse semantic contexts are recognized more rapidly than words encountered in fewer semantic contexts (Adelman et al., 2006; Hoffman et al., 2013; Jones et al., 2012; Johns & Jones, 2022). This effect has been explained in terms of a principle of likely need; words with greater dispersion across semantic contexts are recognized more rapidly because they are useful in diverse contexts (Adelman et al., 2006).

This construct of contextual diversity (CD) has been operationalized in a variety of ways reflecting an increasingly sophisticated understanding of the underlying construct: (a) document range (Adelman et al., 2006), (b) average semantic variation in linguistic co-texts around occurrences of a word (Hoffman et al., 2013), (c) the number of semantically unique occurrences of a word (Jones et al., 2012), and (d) the number of unique discourses or topics in which the word appears (Johns, 2021). Thus, researchers have moved from an understanding of contextual diversity that foregrounds the number of encounters with a word to one that foregrounds the number of unique contexts and discourse situations in which the word is encountered.

It follows then that the number of registers in which a word occurs (its register range) may also be an effective measure of CD. In the text-linguistic tradition, registers have traditionally been defined as varieties of language associated with different situations of use (Biber & Conrad, 2019). Presumably, words that occur in a wide range of registers are also broadly useful.

The present study tests this proposal by using multiple regression on a dataset of 18,000 word types to explain variance in lexical decision task reaction times (LDT RTs) as a function of those words' corpus frequencies, document ranges, and register ranges, derived from 30 registers from the Corpus of Online Registers of English (Biber & Egbert, 2016). Hierarchical model comparisons show that register range explains much of the same variance explained by document range and corpus frequency (roughly 16% above baseline) suggesting that the three metrics all measure the same underlying construct.

These results contradict frequency-based explanations of the processing advantage for words with high CD. Less obviously, however, these findings also contradict explanations focusing on variation in the semantics of words' linguistic co-texts. As some registers have a limited range of semantic environments in which a word may occur (e.g., sports reporting, recipes, religious sermons) and others show significant overlap in semantic environments with other registers (between e.g., personal blogs and travel blogs or encyclopedia articles and historical articles), these results suggest that the key driver of processing advantage in LDT RTs is not encountering a word in diverse semantic contexts, but rather in diverse situations of use, a finding which aligns with studies focusing on unique situational contexts (e.g., Johns & Jones, 2022) and those showing differential processing advantage based on situational variables like task (e.g., Baayen et al., 2016).

#126

Investigating the use of acoustic cues for addressee inference in machine-directed speech

Cassidy Henry, Alayo Tripp, William Idsardi

Humans modulate language behavior based upon the social category of an interlocutor, resulting in distinctive speech registers. Despite the social salience of variation in register, the phenomenon of register use is often not accounted for in end-user spoken dialogue systems. An improved understanding of register is required to support the development of agile speech recognition systems which can both withstand and interpret variation arising from register selection.

To investigate the relationship between perception of speech register and speech addressee, we introduce a two-part experiment including a perceptual and production task. The perception data present a novel benchmark of human capability, while the production data yield a novel parallel corpus of responses directed at humans and machines.

Eighty adult American English speakers (54 female, 26 male) participated. In the perceptual task, participants classified utterances as adult-directed (ADS), machine-directed (MDS), or infant-directed (IDS, as a control). IDS served as a control because it is well-established as perceptually salient and reliably detectable. Stimuli were presented in two acoustic contexts: full-signal speech with complete phonetic detail and low-pass filtered speech (400 Hz cutoff), which preserved intonation and timing while obscuring lexical detail. To obtain speech of varying directedness, it was necessary to sample multiple corpora, as there is currently no publicly available corpus containing parallel productions from individual speakers in these registers. Stimuli were therefore drawn from three corpora:the Buckeye Corpus (ADS) (Pitt et al., 2005), USC Institute for Creative Technologies' Distress Analysis Interview Corpus (MDS) (Gratch, et al., 2014) and CHILDES English Bernstein Ratner Corpus (IDS) (Bernstein Ratner, 1984). From each corpus, 20 utterances were selected after quality filtering, yielding 60 unique items. Each was presented in both full-signal and low-pass filtered form, for a total of 120 stimuli. The lexical similarity, average length, and other attributes of the items were reviewed to ensure that no patterns or similarities could be discerned within a single data set, where possible. Performance was well above chance in both conditions, providing clear evidence that register is encoded in prosodic and temporal features of the speech signal and is perceptually accessible without lexical content.

In the production task, participants produced responses to prompts addressed to a familiar human, an unfamiliar human, and a machine (following Cohn et al., 2024). This task yielded a novel parallel corpus of human- versus machine-directed speech, supporting future computational models of register.

Together, these experiments advancing both theoretical understanding and the development of speech systems robust to socially conditioned variation, providing (i) a perceptual benchmark for register detection and (ii) a new parallel corpus for modeling register in production, to be published in the Institutional Repository at the University of Florida.

#129

Corpus-informed prompt design for LLM-mediated support for researchers with limited proficiency in Academic English

Tony Berber Sardinha, Ana Boconry, Deise Dutra, Walcir Cardoso, Anderson Ávila, Marilisa Shimazumi, Vivian Lameira, Juliana Almeida, Luciana Aguiar de Oliveira, Gabriela Escobar

Writing fluently and appropriately in the academic register presents challenges for researchers with limited proficiency in academic English. Large Language Models (LLMs) can assist in this process, as they have been trained on vast collections of English text and have acquired extensive phraseological and lexicogrammatical patterns (Author, in press; Yamada, 2025). However, LLMs are not register-aware (Author, 2025): they often generate text that appears academic but lacks the full set of linguistic characteristics expected of this register. We have called this process 'register metamorphosis' (Authors), whereby an LLM produces prose that appears scholarly but departs linguistically from the patterns that characterize academic writing. The problem is compounded by the fact that academic register varies across disciplines, with each field displaying its own linguistic configurations (Authors, 2024). To generate discipline-specific research-article prose, an LLM would need to reproduce linguistic features it has not directly learned during training. These features include lexical bundles (Biber, Conrad, & Cortes, 2004; Cortes, 2024), key grammatical features (Authors, 2024), rhetorical moves (Swales, 1990), and underlying parameters of register variation (Biber, 1988, 1995; Authors, 2014, 2019). This paper reports a study that investigated whether LLMs can be guided to produce writing that more closely conforms to the register of research articles through corpus-informed prompts. The prompts conveyed explicit linguistic guidance on key components of academic discourse. First, rhetorical moves were specified, defined as a functional unit within a genre that serves a particular communicative purpose (Swales, 1990). Each move may be realized by several steps, which are recognizable strategies for fulfilling that purpose. Second, lexical bundles were included. These are recurrent multi-word sequences that appear above a frequency threshold across texts within a register (Biber et al., 2004). They play a role in providing academic writing with formulaic fluency and in signaling discourse functions, such as marking stance. Third, the prompts targeted key grammatical features characteristic of research-article prose. These include frequent use of passive voice to downplay agency. Fourth, the prompts incorporated multi-dimensional profiles derived from earlier analyses of register variation. Academic research articles typically show high concentrations of informational density, abstract exposition, and a relative absence of involvement. The specifications for all four components were informed by our previous empirical work and by findings reported in the literature. Through these corpus-informed prompts, we sought to steer a particular LLM (GPT 5) to generate prose that exhibits characteristic phraseology, grammar, rhetorical organization, and overall dimensional tendencies of journal articles. We collected the responses from the prompts and analyzed the presence of the targeted characteristics, comparing them with the same prompts without explicit linguistic guidance to determine the extent to which the LLM was steered to reproduce the expected corpus-based patterns. In this presentation, we will report our findings, with the aim of contributing to the assessment of whether the gap between apparent and genuine academic register in LLM outputs can be narrowed through corpus-informed prompt design.

#131

Affective language and fake news in Brazilian Portuguese

Camila Lívio, Chad Howe

"Fennel tea cures the Coronavirus" and other outlandish headlines flood the social media feeds daily for many Brazilians (Rudnitzki and Scofield 2020). With advances in technology, the internet, and social media, fake news spreads quickly and has been argued to affect democratic processes, as well as lead to tragic consequences (Silva et al., 2020: 1-2). The publication of The Language of Fake News by Grieve and Woodfield (2023) constitutes an important step toward unpacking the linguistic patterns of fake news using corpus data and shedding light on how a better understanding of the stylistic features of these texts is critical to differentiate fake from real news, with the added benefit of promoting media literacy more broadly.

Our proposal seeks to contribute to this body of research by analyzing bigrams of intensifier-adjective in the FakeRecogna corpus (Garcia, Afonso, & Papa, 2022). The corpus contains aligned "true" and "fake news" pairs in Brazilian Portuguese, collected from manually checked fake news on the web (approximately 12,000 news samples). Additionally, it was semi-automatically designed to find corresponding true news for each fake news piece, facilitating a more reliable comparison of the presence or absence of linguistic features. More specifically, this research observes the behavior of intensifiers, defined as elements that scale up the quality of a word, such as in "absolutely amazing" or "totally cool" in English or muito and bem in Portuguese. These affective language structures have been argued to play an important role in the expression of emotions and, importantly, in the detection of fake news (Grieve and Woodfield, 2023: 54), as the use of these particles indicates heightened subjectivity (Anathasiadou, 2017). As a rhetorical device, intensification can be used to emphasize the speaker's views and beliefs in an attempt to establish common ground with their audience and create a strong impression.

To carry out the analysis, we will use a combination of text mining techniques (bigram analysis) and corpus methods (collocation and frequency analysis) to extract bigrams of intensifier-adjectives, with the objective of answering to two main questions: (1) Is there a difference in rate and type between real and fake news in terms of the use of intensifiers? (2) What type of adjectives do these intensifiers modify? Our initial findings suggest that use of intensifiers provides the most effective measure of distinguishing between 'real' and 'fake' news (i.e. compared to adjective use or adjective/intensifier bigrams). With this research, we hope to contribute to the understanding of the language of fake news from an empirical and language-centered standpoint from Latin America, while also discussing issues related to data annotation and the implications for the analysis.

#133

Mapping words to functional clusters with prosodic profiles

Ryan Ka Yau Lai, Lu Liu, Haoran Yan, John DuBois

How words pattern relative to their contextual neighbors has long been recognized to reflect linguistic function (Firth 1957); recent research extends this insight to cooccurrence with constructions and semantic properties (Gries 2010). Here we introduce a prosodic perspective, focusing on how words relate to positions within intonation units (IUs), i.e. "stretch[es] of speech uttered under a single coherent intonation contour" (DuBois et al. 1993:47), which are argued to be a fundamental organizing principle of spoken language (Chafe 1979, 1993, Inbar et al. 2025, Matalon et al. 2025). Prior work found that high-frequency words in the Santa Barbara Corpus (SBC; DuBois et al. 2000-2005) gravitate towards specific IU positions, e.g. IU-initial or IU-penultimate (Lai et al. 2023). These prosodic profiles often correlate with linguistic function, e.g. words that anticipate following noun phrases usually appear near IU endings.

In the current study, we aim to mathematically model the distributional profiles of words within IUs, and extract linguistically interpretable overarching patterns.

We first modeled distributional profiles of the 200 most frequent words in the SBC. Our models captured both IU length variability and position within the IU, counting from either front or back as appropriate. The basic model uses a negative binomial distribution for IU length and a truncated generalized Poisson distribution for position; additional modifications were made for complex (e.g. bimodal) distributions. Models were fitted within a Bayesian framework and selected with Widely Applicable Information Criterion. Posterior predictive checks reveal that these simple probability models accurately portray most words' distributional profiles.

Secondly, to explore how prosodic distribution may reflect syntacto-semantic function, we applied the Mapper algorithm (Singh et al. 2007) to the fitted models, distilling the high-dimensional patterns into a network of nodes. The process involved calculating Jensen-Shannon distances between fitted probability distributions, projected into five dimensions using multidimensional scaling (MDS). The first two MDS dimensions were divided into overlapping subregions; words within each subregion were then clustered using all five dimensions. This created a network where nodes represent clusters and lines indicate shared words between clusters.

We find that each node represents a distributional profile associated with one or more functions. For instance, interjections (e.g., uh, oh) are found in small outlying nodes, concentrated in very short IUs. Other sets of nodes were found to represent nouns and indefinite pronouns which tend to end longer IUs, framing words (e.g., auxiliaries, complement-taking verbs, predicate-level adverbs) concentrated around second position, etc. The non-hierarchical nature of Mapper also helps reveal polyfunctionality. For example, in one node, the words people and things appear in earlier positions to frame generic statements, clustering with nominative pronouns. But in another node, they also refer to more concrete entities, where they pattern with other nouns, including singular person and thing.

We conclude that words' positions within IUs are highly regular, modellable by basic probabilistic models, and contain substantial information on linguistic function. Given the relative ease of obtaining IU transcriptions (Roll et al. 2023, Troiani 2023), they constitute a valuable, cost-effective addition to the corpus semanticist's toolkit.

#137

Applying corpus-assisted discourse studies (CADS) to examine representations of agriculture in The Catholic Worker

Shaya Kraut

The Catholic Worker newspaper is a voice for The Catholic Worker movement, founded in 1933 by Dorothy Day and Peter Maurin. Small-scale agriculture has been a Catholic Worker initiative from the start, with a focus on community and sustainability. This initiative was the idea of Maurin, who came from a French farming family that had lived on and farmed the same land for fifteen hundred years. Inspired by the example of monastic orders such as the Benedictines, he promoted a blend of "cult, culture, and cultivation" for laypeople: religious observance, intellectual and artistic pursuits, and life on the land. The Catholic Worker publishes updates on Catholic Worker farms as well as other articles about agriculture. The newspaper is important societally because of its ability over nearly a century to reconcile traditional Catholicism and radical social justice. This CADS study looks at how the paper frames agriculture in ways that could resonate with readers falling at different points on continua of religiosity, socio-political perspective, and intellectual inclination.

During an interview with a Catholic Worker farmer for a larger study, an interviewee expressed the idea that other living things should be treated according to the philosophy of personalism. Personalism is a key principle of the Catholic Worker movement; it is a spiritual belief in the dignity and worth of each individual person, unrelated to achievement, wealth, or other worldly metrics. She also expressed dissatisfaction with popular culture representations of farm work as menial drudgery, when in fact it can be creative, fulfilling, intellectually stimulating work.

In the spirit of James R. Martin's call for "positive discourse analysis" (PDA), or analysis of advocacy exemplars to complement critical discourse analysis, this study is a corpus-assisted discourse studies (CADS) project looking at the discourse of agriculture in The Catholic Worker over the decade 2012-2021. I hope to learn if and how the paper extends the philosophy of personalism beyond human beings, and how it represents the work of farming. I compiled a corpus of Catholic Worker articles about agriculture using a set of related search terms (e.g., farm, garden, agriculture). Using the application Lancsbox X, I am conducting collocation and concordance analysis based on the terms, following Baker (2006). I am interested in these questions: What relationships are represented between people and the natural world? How is farming represented? Are there distinctive features of the language that other writers dedicated to regenerative agriculture might adopt? Preliminary findings include: Prevalence of first-person pronouns (I), used to share personal reflections and convey lived experiences; farms as hubs of activity involving moving, travelling, and visiting; verbs related to education such as learning at and about farms; and frequency of farmworkers, particularly in expressing solidarity with the United Farm Workers. As I have recently begun analysis, I anticipate many more findings as I continue analyzing the data.

References

Baker, P. (2006). Using corpora in discourse analysis. Continuum.

Martin, J. R. (1999). Grace: The logogenesis of freedom. Discourse Studies, 1(1), 29–56.

#138

Implicit connectives are also eRST signals!

Lin Ai, Amir Zeldes

Discourse relations are inferred connections between propositions which are often underspecified, but can be recovered by speakers. For example, in a sequence such as "Kim Fell. Mary pushed her!" (cf. Asher & Lascarides 2003), we can infer temporal ordering (pushing preceded falling) and causality (pushing caused falling). These relations and others are codified in label inventories used by frameworks such as Rhetorical Structure Theory (RST, Mann & Thompson 1988).

Current work on discourse relation signaling generally distinguishes explicit and implicit relations. In explicit relations, a connective such as 'because' or 'after' can signal causality or temporal ordering, while in implicit ones, such a marker is not present, but can often be reconstructed by speakers ("Kim fell because/after Mary pushed her", cf. Prasad et al. 2018's work on the Penn Discourse Treebank, PDTB).

More recently, work in Enhanced Rhetorical Structure Theory (eRST, Zeldes et al. 2025) has proposed adding a range of signal types to classify the means by which a relation is signaled, including types for connectives like 'because', but also graphical signals (punctuation, layout), lexical signals (e.g. words implying temporality such as 'yesterday') semantic ones (e.g. antonyms to signal contrast), and more. The implementation of this signal taxonomy allows for much more fine-grained distinctions than the presence or absence of a connective as found in PDTB, and allows for the study of signal distributions.

However, a key feature of the PDTB framework remains absent in analyses within the framework of eRST, which does not spell out possible implicit connectives, such as an implied 'because' between the sentences above. In this paper, we propose adding implicit connectives to the existing eRST corpus from Zeldes et al. (2025), by relying on annotations collected in a PDTB analysis of the same underlying corpus data, published by Liu et al. (2024). Using those annotations, we apply two concrete expansions to the data: adding implicit connective signals to relations that are present in both datasets, and adding additional relations corresponding to implicit relations not found in the eRST corpus. We also present a quantitative evaluation of our data quality, as well as some preliminary findings on the distribution of implicit connectives and their distributional relationships with other signaling devices.

References

Asher, Nicholas & Lascarides, Alex (2003). Logics of Conversation. Cambridge University Press.

Liu, Yang Janet, Aoyama, Tatsuya, Scivetti, Wesley, Zhu, Yilun, Behzad, Shabnam, Levine, Lauren Elizabeth, Lin, Jessica, Tiwari, Devika & Zeldes, Amir (2024). GDTB: Genre Diverse Data for English Shallow Discourse Parsing across Modalities, Text Types, and Domains. Proceedings of EMNLP 2025. Miami, 12287–12303.

Mann, William C., & Thompson, Sandra, A. (1988). Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. Text 8(3), 243–281.

Prasad, Rashmi, Webber, Bonnie & Lee, Alan. (2018). Discourse Annotation in the PDTB: The Next Generation. In ACL-ISO Workshop on Interoperable Semantic Annotation, Santa Fe, 87–97.

Zeldes, Amir, Aoyama, Tatsuya, Liu, Yang Janet, Peng, Siyao, Das, Debopam & Gessler, Luke (2025). eRST: A Signaled Graph Theory of Discourse Relations and Organization. Computational Linguistics 51(1):23–72.

#140

Lexical markers of climate discourse in the Polish opinion press: a corpus-based and discourse-historical approach

Dagmara Mateja

This paper examines how lexical markers differentiate climate discourse across Polish opinion weeklies, drawing on a 1.65 million-token corpus of climate-related texts published between 2015 and 2022 (Polityka, Newsweek, Sieci, Wprost, Do Rzeczy). While most corpus-based studies of climate communication have concentrated on Anglophone media, the Polish opinion press, characterized by strong ideological polarization, has received little systematic attention. This study provides the first large-scale corpus-based comparison of climate discourse in this context.

The corpus was purpose-built using machine learning methods for automatic retrieval and classification of climate-related texts, ensuring both coverage and thematic consistency across outlets. Methodologically, the study advances corpus-assisted discourse studies (CADS) by extending keyword-based approaches with statistical rigor and discourse-historical interpretation. Distinctive lexical markers were identified using log-likelihood with Benjamini–Hochberg correction ($q < 0.05$), a procedure that minimizes the inflation of false positives across thousands of tests. Items were retained only if statistically significant and consistently more frequent in the focal outlet. Non-thematic elements such as proper names, paratextual markers, and high-frequency function words were removed through multi-step filtering, and lemmatization inconsistencies were normalized. This procedure yields lexical markers that are both statistically reliable and semantically interpretable.

Integrated into the Discourse-Historical Approach (DHA), these markers serve as entry points into the reconstruction of discursive profiles. The findings reveal sharply differentiated orientations: Do Rzeczy frames climate through geopolitics and sovereignty, Sieci emphasizes economic and sectoral perspectives, Polityka highlights ecological and scientific authority, Newsweek bridges expert discourse and everyday practice, while Wprost embeds climate within entrepreneurial and innovation-oriented narratives.

Together, these results provide a semantic map of climate discourse in the Polish opinion press. The study contributes methodologically by demonstrating how statistically controlled keyword extraction can be integrated with DHA, moving beyond raw keyword lists toward ideologically interpretable discourse profiles. Empirically, it opens the underexplored Central and Eastern European media landscape to systematic corpus-based investigation, showing how climate discourse is linguistically encoded and contested in a region marked by both political polarization and energy transition. This dual contribution illustrates how the combination of machine learning, quantitative corpus methods, and discourse-historical interpretation can advance research at the interface of language, politics, and climate communication.

References

Baker, P. (2006). Using corpora in discourse analysis. Continuum.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B, 57(1), 289–300.

Kilgarriff, A., et al. (2014). The Sketch Engine: ten years on. Lexicography, 1(1), 7–36.

Reisigl, M., & Wodak, R. (2009). The discourse-historical approach. In R. Wodak & M. Meyer (Eds.), Methods of critical discourse analysis (pp. 87–121). Sage.

Stubbs, M. (2010). Three concepts of keywords. Keyword Studies, 3(1), 1–23.

Taylor, C., & Marchi, A. (2018). Corpus approaches to discourse: A critical review. Routledge.

#143

NLP tools for constructing a spoken corpus of endangered language varieties: a case study of the ELIC corpus

Massimo Daul, Austin Jones, John Hale, Margaret Renwick, Zvjezdana Vrzić, Keith Langston

The difficulties of building a corpus of spoken language (Love et al. 2017: 320–321) are compounded when dealing with low- or no-resource language varieties. NLP technologies are essential for overcoming the annotation bottleneck but are challenging to develop with limited training data (Berez-Kroeker et al. 2023: 201–202, Pakray et al. 2025). We report here on the development and evaluation of NLP tools as part of a project to create a corpus of endangered language varieties spoken in the Istria and Kvarner regions of Croatia. The ELIC Corpus contains over 60 hours of sociolinguistic interview data from Čakavian, Fiuman, Istriot, Istro-Romanian, and Istro-Venetian speakers, which will be transcribed and time-aligned at the utterance and word levels and morphosyntactically annotated.

The time required for manual transcription of our interviews ranges from 19-40 hours per hour of audio, which is prohibitive even for a relatively small corpus such as this. Although there are closely related languages for which ASR resources do exist (e.g., Croatian for Čakavian), the performance of models trained on these related languages is not accurate enough to be usable, due to substantial differences in phonology, morphology, and lexicon. Our data also pose additional challenges: the interview format was designed to collect naturalistic data, so there is frequent overlapping of speakers; interviews were generally conducted in the participants' homes, producing recordings of varying quality; and there is significant variation among speakers of the same languages due to differences in local dialects and language contact. To overcome these hurdles, we have fine-tuned Wav2Vec-BERT 2.0 (Meta AI 2024) on languages in our corpus and have developed a speech-to-text pipeline that outputs diarized Praat TextGrid files aligned at the utterance level. Despite small amounts of training data (fewer than 4 hours per variety), the WER is less than 20% in each case. While these automatic transcriptions cannot be used directly, our testing shows that correcting the ASR output can be up to 50% faster than manual transcription from scratch.

The Montreal Forced Aligner (McAuliffe et al. 2017) is used to align transcriptions at the word level. We compare sample forced alignments to manual alignments and describe factors that may affect performance. We have found that different approaches are required in similar use cases. For the diverse Čakavian dialects in the corpus, we achieved acceptable results using MFA's Croatian acoustic model and an adapted version of the Croatian dictionary. For Istro-Venetian and Fiuman, on the other hand, MFA's Italian acoustic model and adapted dictionary performed noticeably worse. We obtained the best results by training new models separately for each of these varieties, despite their similarity (both being colonial Venetian varieties). Our experience also indicates that even when the available training data are very limited, excluding lower quality recordings can improve performance. For our purposes, substantial alignment errors are still frequent enough to require manual correction. While the MFA is a key tool in our workflow, the specific needs of each low-resource variety necessitate careful adaptation.

#146

From “proposing” to “arguing”: academic writing in an English major program

Marine Matte, Simone Sarmento

Although academic writing has been widely examined across genres and disciplines (Nesi & Gardner, 2012; Römer & O’Donnell, 2011), Brazilian undergraduate English Language and Literature programs remain underexplored. This study maps students’ academic writing using a corpus of assignments from English language courses across the program’s eight curricular levels (Matte, 2024). The texts were categorized according to their communicative purposes, following Goulart et al.'s (2022) framework, resulting in eight categories: analyze, argue, explain, give personal advice, narrate, narrate a personal event, propose, and review. For the linguistic analysis, 28 lexico-grammatical features identified as relevant in writing (Biber et al., 2011) were considered. These features are distributed across three groups: clausal (verbs, subordination, conjunctions), phrasal (nouns, adjectives, nominalizations), and intermediate (relative clauses with “that” and wh-forms). The analysis of communicative purposes across the different levels of study shows that in “English 1” texts with the purpose of propose are predominant. In “English 2” and “English 3” there is a diversity of purposes, with argue being the most frequent. In “English 4” texts generally aim to argue and narrate a personal event. In “English 5” and “English 6” argue remains predominant, while in “English 7” and “English 8” narrate a personal event and argue stand out, respectively. The behavior of the linguistic features associated with grammatical complexity in carrying out communicative purposes reveals, among other patterns, that adjectives are more common in review texts, where students describe attributes and offer evaluations; relative clauses with “that” appear frequently in texts with the purpose of explain, which require defining and delimiting concepts in Linguistics and Literature, and nominalizations are common in texts that propose, as this feature is used to name procedures and tasks related to teaching practices. These results highlight the importance of considering different communicative purposes throughout the eight English Language courses and the linguistic resources that realize each purpose. It also demonstrates how corpus-based analyses can capture both situational and linguistic characteristics of academic writing.

#148

Testing a TxTLx approach to variation in dissertation writing

Febriana Lestari

Dissertation writing is a high-stakes academic task that demonstrates graduate students' disciplinary knowledge (Thompson, 2013). Yet writing support that graduate students receive remains limited: (1) few dedicated courses focus on dissertation writing, and (2) graduate courses typically focus on disciplinary content over the language (Becker, 2022). Additionally, this writing is often treated as equivalent to research articles, and published guidelines typically provide only general "self-help" advice (Anderson & Okuda, 2021), with little linguistic or discipline-specific guidance. Research on dissertation writing is also limited in scope, with most studies focus narrowly on macrostructures without linguistic analysis (e.g., macrostructure types in humanities), individual linguistic features (e.g., studies of stance features in the abstract section), specific dissertation sections (e.g., studies on dissertation Introduction section, abstract, etc.). These studies lack a holistic view of dissertation writing as one complete text.

Using a text-linguistic (TxTLx) approach to register variation (Biber et al., 2020, 2021), the present study examined situational and textual characteristics as well as the co-occurrence of linguistic features at the text level. The study was carried out in three stages: (1) analysis of situational and textual characteristics; (2) corpus design and construction of Dissertation Register Corpus (DRC); and (3) linguistic analysis followed by functional interpretation. The situational and textual analysis was applied across four disciplines (applied linguistics, psychology, engineering, and veterinary medicine) using the DRC, consisting of 172 dissertations. The linguistic analysis piloted additive MDA (Sardinha et al., 2021) on the applied linguistics sub-corpus (43 texts; 2,617,973 words), drawing on Gray's (2011; 2015) four dimensions (e.g., D1: Academic Involvement and Elaboration vs. Informational Density; D2: Contextualized Narration vs. Procedural Discourse).

Preliminary findings identify four dissertation sub-registers: (1) monograph, (2) hybrid monograph, (3) article-based, and (4) hybrid article-based. Each discipline demonstrates distinct trends (e.g., applied linguistics is dominated by the monograph; psychology is largely monograph-based but shows movement toward the hybrid monograph and hybrid article-based). Findings from additive MDA suggest that dissertation writing is distinct from research article writing, even when article-based dissertations include published or publishable research articles. Overall, the linguistic analysis shows distinct functional variation across dissertation sub-registers. For instance, in Dimension 1, all sub-registers in applied linguistics show greater academic involvement than informational density, with the article-based type being the most informationally dense. This study highlights implications for pedagogy, including the need for dissertation-specific courses and materials that reflect disciplinary variation, and for future research that accounts for both situational contexts and linguistic variation.

#149

Comparing reporting verb use across L2 student writing and applied linguistics articles: a replication study

Duong Nguyen, Men Truong

Academic writing often relies on reporting verbs (RVs) to incorporate prior research and strengthen arguments. Previous studies have examined RV use in L1 and L2 student writing (e.g., Charles, 2006a; Kwon et al., 2018) and in published research articles. However, research on published writing has mostly focused on literature review sections or small, single-journal corpora. Less is known about how experienced writers use RVs across article sections and disciplinary subfields.

This study investigates reporting verbs (RVs) use in the Applied Linguistics Research Article Corpus (Gray et al., 2014), which consists of 150 Applied Linguistics research articles (1.1 million words) representing five sub-disciplines. Replicating the methodology of Kwon et al. (2018), 53 RV lemmas categorized into four semantic groups — Argue, Show, Find, and Think (Charles, 2006a) — were extracted using AntConc 3.5.9 (Anthony, 2020). Both raw and normalized frequency per 100,000 words were then calculated. Additionally, the verbs' rhetorical functions — textual reference, self-reference, and uncited generalization (Kwon et al., 2018) were manually coded, and the percentage of use within each rhetorical function category for each semantic category were also computed.

The results of the current study were compared with the findings in the original studies to highlight the difference between L2 student writings and Applied Linguists' writings. It was shown that Applied Linguistics researchers employed RVs less frequently than L2 students (646.61 per 100,000 words). Argue verbs were the most frequent, followed by Show, Find, and Think. While AL researchers rarely used Think verbs, L2 students relied heavily on them. The most frequent RVs were suggest, show, and find. RVs were primarily used for self-reference, particularly with Show and Argue verbs in combination with non-human subjects, it-clefts, and passive constructions. This pattern reflects the rhetorical stance of AL writing, where authors foreground their own findings while maintaining an impersonal tone—contrasting with students' reliance on textual reference in literature reviews.

The comparison with Kwon et al.'s (2018) findings highlights two factors that possibly shape the use of reporting verbs: register and writers' experience with academic discourse. Pedagogical implications include teaching the strategic use of Show and Find verbs, along with syntactic patterns such as non-human subjects, it-clefts, and passives to support objective stance-taking. Instruction should also address how semantic categories, rhetorical functions, and syntactic structures interact in effective academic writing.

References

Anthony, L. (2020). AntConc (Version 3.5.9) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software.

Charles, M. (2006a). Phraseological patterns in reporting clauses used in citation: A corpus-based study of theses in two disciplines. English for Specific Purposes, 25(3), 310–331. https://doi.org/10.1016/j.esp.2005.05.003.

Gray, B., Egbert, J., & Qian, M. (2014, September). Internal representativeness and specialized corpora: The influence of topic on the stability of linguistic findings in a disciplinary writing corpus. Paper presented at the American Association of Corpus Linguistics Conference. Flagstaff, AZ.

Kwon, M. H., Staples, S., & Partridge, R. S. (2018). Source work in the first-year L2 writing classroom: Undergraduate L2 writers' use of reporting verbs. Journal of English for Academic Purposes, 34, 86–96. https://doi.org/10.1016/j.jeap.2018.04.001.

#156

CRetor: an annotated corpus of rhetorical strategies in Mandarin counter speech

Xiaoyu Chen, Chenfeng Su, Michael Bennie

CRetor introduces the first annotated corpus of rhetorical strategies in Mandarin counter speech. We assemble materials from a large pool of user generated content; specifically more than 800 questions collected from Zhihu and Baike. Each question is evaluated with three choices yes; no; potentially to identify whether it invites or embodies pejorative meaning that calls for counter speech. Building on this collection we expand existing taxonomies so that they capture culturally distinct ways Chinese speakers confront stereotypes in online discussion. The taxonomy is designed to be readable by researchers and practitioners; it names strategies in clear terms and aligns them with examples that can support model training and evaluation.

Basic indicators derived from a focused primary analysis show clear quantitative patterns. Our analysis of 72 questions reveals a balanced dataset; 54.2 percent potentially pejorative versus 45.8 percent none. We observe a large difference in text length; potentially pejorative questions are much shorter with an average of 43 characters while non pejorative questions average 92 characters. A Point Biserial correlation test confirms a moderate relationship between the non pejorative label and increased text length with $r = 0.31$ and $p < 0.01$; hostile inquiries tend to be brief while informational questions require more text. Thematic analysis complements these findings. Potentially pejorative questions often use challenging language such as why or is it not and concentrate on sensitive topics such as gender and nationality; non pejorative questions cluster around neutral topics and request background or definitional information.

We then annotate counter speech responses to these prompts and identify strategy preferences that diverge from patterns reported in English language corpora. Chinese counter speech is typically more concise; it prioritizes structural explanations labeled EXTERNAL FACTORS such as pointing to policy; institutions; or historical context; or it relies on factual refutations labeled COUNTEREXAMPLES that present concrete evidence against a generalization. In contrast English language datasets report more frequent use of shifting focus to a group's other positive attributes labeled ALTERNATIVE QUALITIES. We also note pragmatic features that shape the delivery of Mandarin counter speech; frequent use of discourse particles to soften stance; early placement of evidence before evaluation; and explicit acknowledgment of uncertainty when information is incomplete. These observations suggest that message length and information structure interact with strategic choice; short hostile prompts tend to elicit compact replies that present facts first; longer neutral prompts invite extended explanations that build shared context. As such, the dataset enables controlled comparisons between human and machine produced counter speech in Mandarin and in other languages; it is therefore a foundation for cross linguistic transfer studies and for robust safety benchmarks.

#157

Register-based trends in the grammatical complexity of student writing

Bethany Gray, Duong Nguyen, Febriana Lestari, Kimberly Becker

Based on the distinctive register-based patterns in the use of grammatical complexity features documented by decades of corpus research (Biber et al., 1999; Biber & Gray 2016), Biber et al. (2011) proposed a series of developmental stages in which novice writers are hypothesized to produce fewer clausal and more phrasal complexity features as they develop. A number of studies have shown general support for this developmental trend in student writing (e.g., Staples et al. 2022, 2016; Lan et al. 2022, Larsson et al. 2024). At the same time, these studies have not yet observed full alignment with the density of such features in published academic writing. Researchers have hypothesized that this lack of convergence with published norms may be due to register (e.g., that the student texts being analyzed were not the highly-informational, research-based texts that make up most academic writing corpora), level (as primarily undergraduate and first-year graduate students were being analyzed, whereas published academic writing is most often produced by professional academics), or both.

The present study addresses these issues by investigating a fuller range of student writing in the disciplines of linguistics and engineering. That is, the study analyzes the patterns of use for features associated with grammatical complexity across three contexts:

- undergraduate students writing for their discipline-specific courses (332 texts representing 5 registers, sampled from MICUSP (Römer & O'Donnell, 2004) and BAWE (Nesi et al., 2004-));
- early-career students writing for the purposes of their graduate coursework (927 texts representing 7 specific registers, drawn from CorGrad (Becker, 2022); and
- advanced graduate students completing doctoral dissertations (86 texts representing four major 'macrostructures' or dissertation types from the Dissertation Register Corpus (DRC; Lestari, forthcoming).

The corpora were tagged using the Biber tagger and the Developmental Complexity Tagger (Gray et al. 2019), which adds tagfields for 24 clausal and phrasal complexity features from the developmental stages in Biber et al. (2011). Comparisons of mean rates of occurrence across level (UG coursework vs. Grad coursework vs. Grad dissertations), discipline, and sub-registers show systematic patterns of use across all three factors. However, the clearest patterns appear to be primarily associated with register, rather than level or discipline. In the presentation, we explore these linguistic patterns, associating the varying patterns of use for phrasal and clausal complexity features along the hypothesized developmental stages with the situational characteristics of the type of writing produced by students at each stage. We end with implications for graduate coursework development.

#159

Framing fertilization:
a corpus analysis of gamete-centered language and metaphor in Japanese

Lauren Polak, Kaori Idemaru, Cindi SturtzSreetharan

Threats to female bodily autonomy are increasingly visible worldwide, and one of the factors that has the most profound influence on this situation is the language used to describe female bodies and how their functions are explained. Foundational work by Martin (1991) examined the ways in which the fertilization process was described in English, particularly in the discourse of medical experts. Her research revealed anthropomorphized and androcentric metaphors were common, with sperm being framed as heroic agents rescuing passive eggs from the horrific fate of "waste". She connected this fairytale metaphor to the gender roles commonly assumed by those who carry a particular gamete (i.e. egg = woman = passive). More recent studies (e.g., Almeling, 2023) show that metaphors such as this one still persist in the general public's understanding of fertilization, albeit with some shifts to a more equal-actors metaphor among more gender-egalitarian individuals. However, previous investigations have been limited to English and a mostly North American context. This leaves a gap in how we understand reproduction and use language to describe it in a multilingual, multicultural way.

The current study addresses this gap by examining Japanese-language reproductive discourse using corpus linguistic methods. We ask (1) how gametes are linguistically framed in Japanese reproductive discourse; (2) what metaphors and collocational patterns emerge, and how they compare to previous English findings; and (3) how effective are corpus methods at revealing broader cultural ideologies embedded in language?

Our data consists of both pre-existing corpora and self-compiled corpora. We use collocation data to identify what kinds of adjectives and verbs are used to describe gametes and their movements as well as the constructions and metaphors in which they are placed. We also use concordance data to reveal the broader semantic contexts in which these words appear. Additionally, keyword analysis is used ascertain whether certain concepts are uniquely central to explanations of the fertilization process, especially upon cross-linguistic comparison.

Initial analyses of our self-compiled high school biology textbook corpus have already revealed promising collocational data, with sperm being used more in agentive, active sentence and constructions, while eggs are framed more passively. Concordance data has shown a lack of "fairytale" framing that seems to be prevalent in English. Preliminary findings suggest that while Japanese discourse may avoid overt metaphors—especially as romanticized as those present in the English-language discourse—there is still a slight gendered asymmetry. This study in progress builds on a growing body of work showing how language and metaphor both influence and are influenced by societally prevalent ideologies. Additionally, we aim to demonstrate the versatility of corpus methods by showing their utility in addressing even questions that may be considered outside the corpus linguist's typical purview.

#161

College entrance written exam analysis from a multidimensional perspective: a corpus linguistics approach

Juliana Barreto

The research reported here intends to analyze the evaluation of written texts in the college entrance essays produced by undergraduate applicants. More specifically, this study verifies the relationship between the composition tests, written by applicants during the admission process, and the varying dimensions of Brazilian Portuguese, as presented in Berber Sardinha, Kauffmann, and Acunzo (2014). The research uses the theoretical framework of Corpus Linguistics and the methodological approach of Multidimensional Analysis. The study corpus is composed of one hundred essays written by applicants for admission to undergraduate courses in higher education, tagged by Palavras parser and post-processed with a script that calculates the score of each text according to all six variation dimensions of Brazilian Portuguese. Initially, it is assessed how the applicants' texts relate to the six Brazilian Portuguese dimensions. Then, the variation in relation to the grades awarded to these essays by examiners is observed in order to determine whether and which correction criteria were met, based on the scores of Brazilian Portuguese multidimensional analysis. Hence, the outcomes here are likely to provide important contributions to the field of textual production in Portuguese in Brazil, considering that it is vital to develop a more accurate understanding of the language in use applied to the teaching and learning of argumentative text production, written by applicants during their admission to undergraduate courses in Higher Education.

#162

From syntax to discourse:
LLM-assisted annotation of code-switching in typologically diverse language pairs

Olga Kellert

In this talk, I will present ongoing research on leveraging large language models (LLMs) to annotate discourse-pragmatic features in bilingual corpora of code-switching. The study builds directly on our previous work, where we successfully applied LLMs for syntactic annotation of Spanish-Guaraní and Spanish-English data using the Universal Dependencies framework. While syntactic annotation has provided valuable insights into structural switch points, the current project aims to move beyond syntax and investigate how code-switching interacts with discourse functions. This perspective allows us to ask whether switches are random or whether they systematically align with pragmatic purposes in conversation.

We focus on two corpora that exemplify typologically different language pairs: Spanish–English, which involves two global languages with long histories of bilingualism in communities such as Miami, and Spanish-Guaraní, which involves an indigenous and a colonial language with distinct sociolinguistic profiles in Paraguay. The contrast between these cases allows us to test whether LLM-driven annotation can generalize across different linguistic and cultural contexts, and whether discourse-level motivations for code-switching differ between global and local bilingual settings.

The methodological innovation of our study lies in extending LLM-based annotation pipelines to discourse features. We designed a set of annotation prompts guiding the LLM to identify functions such as topic shift, emphasis, repair, quotation, or addressee orientation at the point of a switch. These functions are based on established frameworks in discourse pragmatics, adapted to bilingual data. Because LLMs can capture contextual meaning beyond syntax, they offer a unique opportunity to scale annotation of discourse features in ways that would be prohibitively time-consuming with traditional manual methods.

To ensure reliability, we conduct systematic evaluation against human gold annotations. Annotators with expertise in discourse analysis and bilingual corpora independently label subsets of the data, allowing us to assess agreement between human and model outputs. Preliminary results suggest that LLMs achieve reasonable accuracy in capturing discourse motivations, though performance varies across functions. For example, the models perform better at identifying repairs and quotations, which involve explicit markers in the text, but struggle with subtler functions such as emphasis or stance. Interestingly, the Spanish-Guaraní data present additional challenges due to morphological complexity and the presence of under-documented pragmatic markers.

The broader goal of this research is twofold. First, we aim to enrich the theoretical study of code-switching by integrating discourse-pragmatic dimensions into large-scale corpus analysis. This allows us to answer questions such as: Do speakers switch languages more often when signaling topic shifts? Are repairs preferentially marked by one language? Does the social status of a language influence the discourse functions where it appears? Second, we aim to test the extent to which LLMs, trained predominantly on monolingual and high-resource data, can adapt to low-resource, bilingual, and pragmatically complex contexts.

In conclusion, our study demonstrates the potential of LLMs as tools for annotating not just syntactic but also discourse-level features in bilingual corpora. While human validation remains crucial, LLM-assisted annotation offers a scalable pathway to investigating the pragmatic dimensions of code-switching across typologically diverse language pairs.

#163

Not just "*may*" and "*might*": mapping multi-word hedges in research articles

Wesley Acorinti, Alexander Holmberg

Research shows that research articles use hedges (Hyland, 1998) and formulaic language (Conrad & Biber, 2005), which varies across article sections (Cortes, 2013) and disciplines (Hyland, 2008). While hedging (i.e., expressing caution or uncertainty) is well-studied, most studies focus on single-word hedges (Hyland, 1995; Wang, 2022), leaving the role of lexical bundles (i.e., recurrent multi-word sequences) underexplored. In contrast to single-word hedges, investigating hedging lexical bundles (HLBs) seems particularly relevant, as bundles are often used to structure discourse (Biber & Barbieri, 2007), signal disciplinary competence (Hyland, 2008), and may assume hedging functions even when composed of forms not typically regarded as hedges, such as "always" in "it is not always" (Chen & Baker, 2010).

This corpus-based study examines Hedging Lexical Bundles (HLBs) in humanities research using the 7.3-million-word Corpus of Humanities (COURHUM) to address the following research questions: (1) What is the proportion of HLBs across sections and disciplines? (2) Which HLBs are discipline and section-specific?

Five-word bundles were extracted from eight humanities and social science disciplines and four sections using frequency (min=10) and dispersion (min=5) thresholds. HLBs were then identified by drawing from hedges identified by prior studies (e.g., Aull, 2015; Demir, 2018). Additionally, LBs in COREHUM were manually checked to prevent overlooking potential HLBs.

Of 1,128 bundles, 143 (~12.5%) serve a hedging function (e.g., "may be due to the"). HLBs are most frequent in introductions and results/discussion sections, and least common in methodology sections, where uncertainty may be less common. Some HLBs are domain general, while others are discipline- (e.g., "this study suggests that" in social sciences) or section-specific (e.g., "can be interpreted as" in results/discussion sections), indicating that certain disciplines and sections appear to have preferred HLBs.

Findings can be used to inform English for research publication purposes, where mastering lexical bundles supports participation in academic communities (Hyland, 2008).