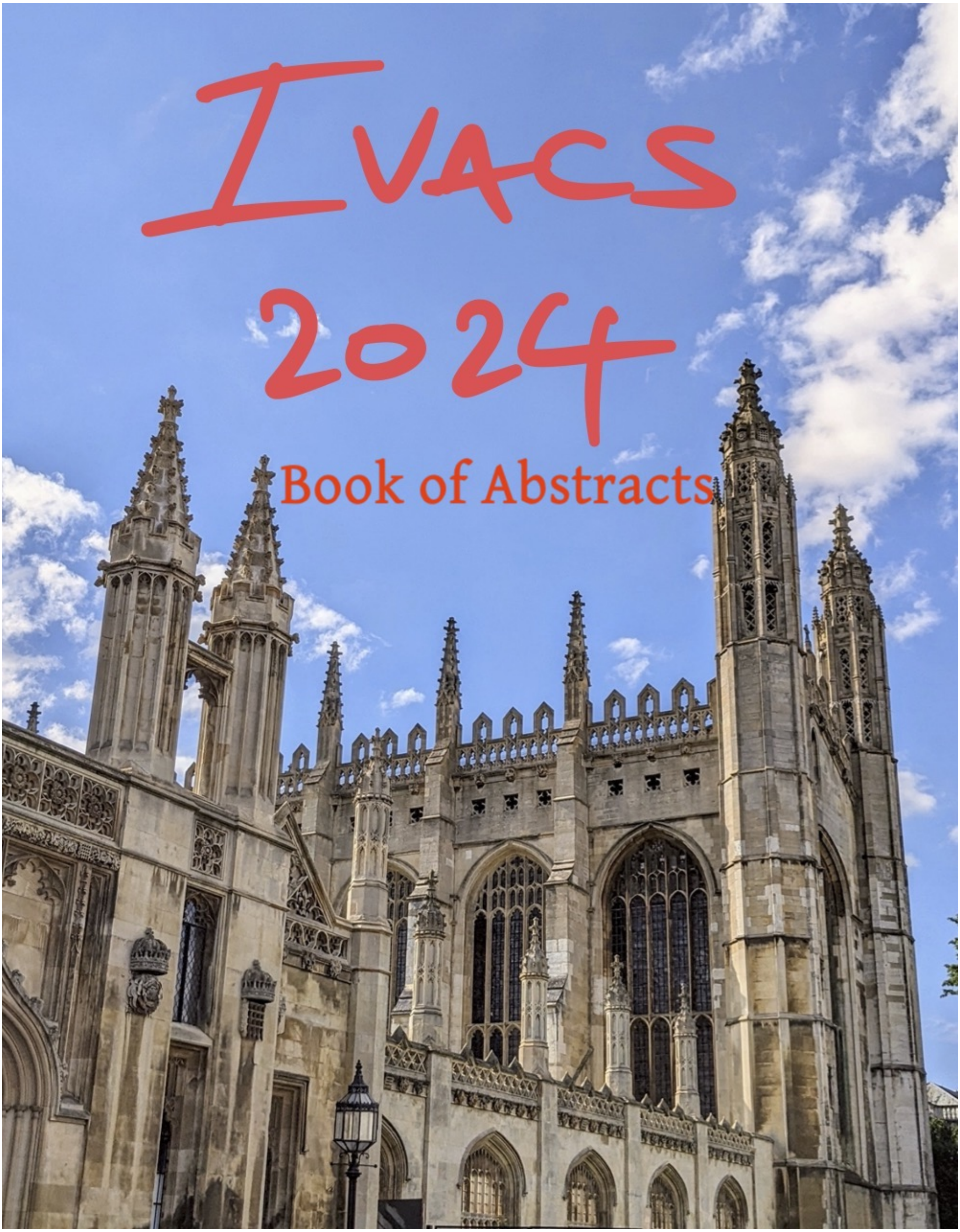
IVACS 2024

Book of Abstracts



Contents

Welcome

Welcome, everyone, to Cambridge and the 11th Inter-Varietal Applied Corpus Studies Biennial Conference.

We are very pleased to be hosting you here in the Crausaz Wordsworth Building at Robinson College. We hope you enjoy the conference, connecting with friends and colleagues old and new, while sampling some of the 92 talks and posters in the programme.

The conference would not be possible without the hard work of our organising committee, advisory board, and team of reviewers. We thank you! Thanks also to our volunteer team who are on hand to help you at the venue, the session chairs, and to the conference team at Robinson who have supported us so fully.

The IVACS 2024 Organisers: Andrew Caines, Anne O'Keeffe, Paula Buttery, Christopher Fitzgerald, Gabrielle Gaudeau.

Advisory Committee: Fiona Farr, Eric Friginal, Dawn Knight, Pascual Pérez-Paredes, Elaine Vaughan.

Volunteer Team: Diana Galvan-Sosa, Yuan Gao, Gabrielle Gaudeau, Kata Szabo.

Schedule Overview

All times are BST (UTC+1).

Day 1: Tuesday 16th July
0900 Welcome, Registration, Coffee
0930 Keynote 1
1030 Break
1045 Talks 1
1145 Break
1200 Talks 2
1300 Lunch
1400 Talks 3
1500 Break
1530 Panel Discussion
1630 Poster Session & Drinks Reception
1830 Conference Dinner

Day 2: Wednesday 17th July
0900 Welcome, Registration, Coffee
0930 Keynote 2
1030 Break
1045 Talks 4
1145 Break
1200 Talks 5
1300 Lunch
1400 Talks 6
1500 Break
1530 Talks 7
1630 Conference Close

Day 1: Tuesday 16th July

0900 - Welcome, Registration, Coffee

In the foyer of the Crausaz Wordsworth Building, Robinson College, Cambridge.

0930 - Keynote 1, Plenary room

Dr Brian Clancy, Mary Immaculate College, Ireland

Bio: Brian Clancy is a lecturer in applied linguistics at Mary Immaculate College, Ireland. His research work focusses on the blend of a corpus linguistic methodology with the discourse analytic approaches of pragmatics and sociolinguistics. His primary methodological interests relate to the use of corpora in the study of language varieties and the construction and analysis of small corpora. His published work explores language use in intimate settings, such as between family and close friends, and the language variety Irish English. He is author of *Investigating Intimate Discourse: Exploring the spoken interaction of families, couples and close friends* (Routledge, 2016) and co-author, with Anne O'Keeffe and Svenja Adolphs, of *Introducing Pragmatics in Use* (Routledge, 2011 and 2020).

Title: **A Many-Splendored Thing: (Corpus) Perspectives on Intimate Discourse**

Abstract: This paper takes a pluralistic, corpus approach to explorations of intimate discourse. Intimate discourse, that is to say the conversation between couples, family and close friends, is an integral part of our lives and everyday spoken world. Much of the initial literature on intimacy explored what might be referred to as 'established intimacy', with a focus on parents with children or friends that lived in shared accommodation over a long period of time (see Clancy, 2016). One of the reasons for the privileging of these sites was the issue of the accessibility of the context, which meant that much of the data collection was, by necessity, opportunistic. However, alongside today's growing proliferation of data in the public domain, and more openness to seeing the potential in different types of linguistic data, there are more routes to intimate discourse than ever before. This has allowed researchers to continue their focus on established intimacy, while also investigating what we can term 'nascent intimacy', such as initial encounters (see, for example, Haugh and Sinkeviciute, 2021), and indeed, initial romantic encounters. Using a comparative corpus approach that primarily utilises first-generation concordancing tools such as keyword lists and concordance lines, and a range of small corpora collected to represent spoken Irish English, this paper seeks to marry findings from both established and nascent intimacy in order to determine how they might inform and complement one another. In doing so, it argues for the value of applied corpus approaches, a 'first principles' orientation to seeing into datasets, and how intimate discourse, the delicate calibration of language that is core to our interpersonal relationships, should be highlighted and valorised.

References:

Clancy, B. 2016. Investigating Intimate Discourse: Exploring the Spoken Interaction of Families, Couples and Friends. London: Routledge.

Haugh, M and Sinkeviciute, V. 2021. 'The pragmatics of initial interactions: Cross-cultural and intercultural perspectives.' Journal of Pragmatics, 185, 35-39.

Chair: Prof Michael McCarthy

1030 - Break, Foyer

1045 - Talks 1, Plenary room

Chair: Odette Vassallo

- 10:45 **Encarnacion Hidalgo-Tenorio, Miguel Angel Benitez-Castro, Aritz Gorostiza, Juan Luis Castro-Peña** : Nutcracker, a semi-supervised algorithm for the detection of online extremism and disinformation

Abstract

Introduction: This paper aims to elucidate the motivations, rationale, and methodological foundations underlying the collaborative efforts between linguistics and computer science that have led to the development of Nutcracker, a cutting-edge semi-supervised algorithm designed for detecting extremist profiles, radicalisation and disinformation on social networking sites. This algorithm, accessible at https//nutcracker.ugr.es, is a result of a transdisciplinary collaboration initiated under the auspices of five national and European research projects, all spearheaded by the University of Granada since 2017.

Methodology: The fruitful synergies established between linguists and computer scientists have progressively empowered the training and implementation of Nutcracker. The algorithm, rooted in what we call deep-relations, focuses on the following objectives: (1) differentiating extremist from non-extremist profiles; (2) distinguishing potential purveyors of disinformation from those who are not involved in disseminating misleading content; (3) identifying user networks through linguistic patterns reflecting attitudes and sentiments (as exemplified in Francisco & Castro, 2020; Francisco, Benítez-Castro, Hidalgo-Tenorio & Castro, 2022).

Key Components: Nutcracker relies on various ontologies, meticulously identified by the discourse analysts in the team (Dhiab-Hassan, Benítez-Castro & Hidalgo-Tenorio, 2018). The algorithm's ability to detect attitudes and sentiments is built upon an ontology derived from Systemic Functional Linguistics' Appraisal Theory. Specifically, it incorporates a psychologically-inspired version of the taxonomy proposed in Benítez-Castro & Hidalgo-Tenorio (2019).

Conclusion: The collaborative endeavours between these disciplines have culminated in the creation of Nutcracker. By seamlessly integrating insights from linguistic analysis and cutting-edge computational techniques, this algorithm serves as a compelling demonstration of the effectiveness of transdisciplinary research in addressing complex challenges at the intersection of technology and societal issues.

- 11:05 **Theodora Alexopoulou** : The influence of L1 typology on the acquisition of the L2 English article: a large scale corpus study

Abstract

Large learner corpora and data from L2 educational settings (exams, teaching) provide data from learners with a variety of L1 backgrounds and, have, thus, enabled investigations of the effect of the L1-L2/Ln typological similarity/linguistic distance on L2/Ln outcomes that are usually not available through small scale lab-based studies. Schepens et al 2020 show that linguistic distance between L1 and Ln Dutch, measured through a combination of morphological, lexical and phonological measures, predicts proficiency scores in the L2/L3 Dutch STEX exams while L1-L2 linguistic proximity mediates age-related decline in language learning learning (Schepens et al 2022). Murakami and Alexopoulou 2016 show that the availability of a congruent morpheme in L1 (e.g. article, tense marking etc.) predicts accuracy in Cambridge Assessment English proficiency exams. One crucial question is whether the acquisition of individual features (e.g. articles) depends solely on the availability of a congruent element in the L1 or whether, in addition, broader typological differences between L1 and L2/Ln guide learners in approaching their input, influencing their learning. In other words, whether the effect of L1-Ln linguistic distance on broad outcomes (Schepens et al 2020) arises as an aggregate of similarities of individual features or whether overall typological similarity impacts on the acquisition of individual features.

We present a corpus investigation considering item-level typological similarity in terms of the availability of an article in the L1 and broader typological similarity in terms of the linguistic distance between L1 and L2 captured through a variety of lexical, morphosyntactic and phonological measures of linguistic distance. We analyse the accuracy of the use of the definite and indefinite English articles in around 0.5 million writings from EFCAMDAT from learners with eleven typologically diverse L1s. Our results indicate that L1 influence arises from the combination of item level L1-L2 differences, that is in the availability of an article in the L1, as well as broader properties of the L1 grammar. In particular, lexical and syntactic distance are predictors of L2 accuracy in the use of the articles. Further, we find that the availability of a definite article in L1 predicts article omission for both definite and indefinite articles, while the availability of an indefinite article is not predictive of article omission rates.

- 11:25 **Elaine Riordan, Fiona Farr, Andrew Caines, Paula Buttery** : The Teacher-Student Chatroom Corpus: Exploring Student Teachers' Linguistic Choices and Perceptions

Abstract

Facilitating various types of interactions between student teachers and their language students are core elements of professional development and training in language teacher education programmes (Farr et al, 2019). Teaching practice is one of these types of interactions, and since the COVID-19 pandemic, online communication is increasingly being used in this microteaching context (Sezaki et al, 2023). One particular type of online communication our student teachers used for part of their teaching practice component of their programme was a chatroom, a synchronous, text-based environment which has an established history of use in language learning and teaching (Burnett, 2003; Chen et al, 2009).

This paper reports on data from the teacher-student chatroom corpus (Caines et al 2020), focussing on one-to-one teaching sessions between student teachers on an MA in TESOL and their English language students. These sessions were held in week 6 and 9 of a 12-week teaching semester for 30 minutes each. Perceptive data from two focus groups are explored, one after the student teachers had conducted their teaching sessions and the other after the student teachers had examined their chatroom session transcripts. Recurring themes and insights gleaned from their experiences are used to inform the ensuing corpus analysis. Linguistic features of their discourse are then explored using corpus-based techniques, including for example, elicitation and questioning techniques, levels of formality and giving instructions. These are also compared over both sessions to track possible changes or developments. The findings shed light on the approaches these student teachers use to make meaning in this type of CMC, how language can be used in place of visual cues, and what this environment offers novice language teachers.

Burnett, C. (2003) 'Learning to chat: Tutor participation in synchronous online chat', Teaching in Higher Education, 8(2), 247-261.

Caines, A., Yannakoudakis, H., Edmondson, H., Allen, H., Perez-Paredes, P., Byrne, B. and Buttery, P. (2020). 'The Teacher-Student Chatroom Corpus'. 9th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2020). https://doi.org/10.3384/ecp2017510

Chen, Y., Chen, N. S. and Tsai, C. C. (2009) 'The use of online synchronous discussion for web-based professional development for teachers', Computers and Education, 53, 1155-1166.

Farr, F., Farrell, A. and Riordan, E. (2019). Social Interaction in Language Teacher Education. EUP.

Sezaki, H. , Lei, Y., Xu, Y., Hachisuka, S., Warisawa, S., and Kurita. K. (2023) 'Online Technology-Based Microteaching in Teacher Education: A Systematic Literature Review'. Procedia Computer Science, 225: 2487-2496.

1045 - Talks 1, Syndicate 1

Chair: Brian Clancy

- 10:45 **Michael T. L. Pace-Sigge** : Deborah, Linguist yet Professor Michael. How British corpora reflect gender-relation through forms of address

Abstract

This paper investigates the use of terms of address in relation to a number of female and male names. As research as early as the 1990s (see, amongst others, Acker, 1990; Lakoff and Lakoff, 1990; Tannen, 1994; Wodak, 1996) has shown, there is a clear link in the discourse between work roles and gender. More recently, large data-sets allow research into the actual, natural language usage that highlights in how far females are 'automatically' named second after males, e.g. Wright and colleagues (2005) who consider frequency distribution. This corpus-assisted research targets British English use as recorded in the BNC 1994 and BNC 2014 to provide a basis for a qualitative and diachronic look at the uses of such terms like professor, director, minister etc. as node words. This allows to observe whether there have been any changes over two decades. The use of these corpora provides an empirical snapshot of the choice of address employed in Britain in the 1980s/1990s and then in the early 2000s - in particular in newsprint of the time. It also gives insight in the concrete nesting (cf. Hoey, 2005) of female names as compared to male names in mainstream discourse; this enables the construction of how the readership is psychologically primed to connect positions of responsibility and learning with the idea of 'maleness'. Even small differences - like, for example, whether a title or a name are given first - indicate subliminal differences. Whereas no single area of life (academia, business, politics) gives clear parity for females and males according to this data, there is, nevertheless, clear evidence of marked progress. However, while UK politics seems to have shown the greatest move towards parity, overall, the changes are uneven and there are, in fact, areas where fewer females seem to appear in 2014 compared to 1994.

References

British National Corpus 1994: http://www.natcorp.ox.ac.uk/ (last accessed 21/10/2023).

British National Corpus 2014: User Manual and Reference Guide, Version 1.1. (BNC2014). http://corpora.lancs.ac.uk/bnc2014/doc/BNC2014manual.pdf (last accessed 21/10/2023).

Hoey, M. (2005). Lexical Priming. A new theory of words and Language. London: Routledge

Lakoff, R. (2003). Language, Gender, and Politics: Putting "women" and "power" in the same sentence. Holmes, J. & Meyerhoff, M (eds.) The Handbook of Language and Gender, 161-178.

Lakoff, R. T., & Lakoff, R. (1990). Talking Power. Basic Books.

Mollin, S. (2012). Revisiting Binomial Order in English: Ordering Constraints and Reversibility. English Language and Linguistics, 16(1), 81-103.

Motschenbacher, H. (2013). Gentlemen Before Ladies? A Corpus-based Study of Conjunct Order in Personal Binomials. Journal of English Linguistics, 41(3), 212-242.

Tannen, D. (1994). Talking from 9 to 5: How women's and men's conversational styles affect who gets heard, who gets credit, and what gets done at work. New York: William Morrow and Company.

Wodak, Ruth (1996) Power, Discourse, and Styles of Female Leadership in School Committee Meetings. In: Discourse and Power in Educational Organizations. Cresskill, N. J., Hampton Press, pp. 31-54.

Wright, S. K., Hay, J., & Bent, T. (2005). Ladies First? Phonology, Frequency, and the Naming Conspiracy. Linguistics, 43(3), 531-561.

- 11:05 **Ninuk Krismanti, Liam Murray, Brona Murphy** : A corpus-based study on an Indonesian University EFL Teacher Translanguaging Strategies

Abstract

Given the context of Indonesia as a multilingual country, translanguaging is a common feature of daily interactions. In EFL classrooms, the use of English as a medium of instruction adds another layer to the language practices of Indonesian teachers and students. However, despite being an integral part of classroom discourse, translanguaging remains underexplored (García, 2009; Wei, 2018). In Indonesia, the existing studies are narrow in scope as they only focus on attitudes and perceptions (Rasman, 2018; Khairunnisa and Lukmana, 2020; Krismanti, 2022; Raja et.al., 2022). Furthermore, approaching translanguaging in the classroom using corpus study has been rarely done (Gilquin, 2022). Therefore, this study is set to explore translanguaging strategies used by an Indonesian EFL teacher in the university context. This research aims at answering two questions: 1) to what extent translanguaging is manifested in the classroom and 2) what are the roles of translanguaging?. The data for this study is drawn from a pilot study taking place in a university located in Banjarmasin, Indonesia. In this area, there are three languages involved: Bahasa Banjar as the local language, Bahasa Indonesia as the national language, and English as the target language. The corpus built for this research is expected to give a description on how the teacher alternates between languages and for what purposes they are doing that. This research is expected to contribute to the emerging translanguaging study in the context of a multilingual nation where translanguaging is not driven by immigration as the case in USA, UK, and some other parts of the world (Emilia and Hamied, 2022; García, Johnson, and Seltzer, 2016; Garcia et.al., 2023).

References:

Emilia, E., and Hamied, F. A. (2022) 'Translanguaging Practices in a Tertiary EFL Context in Indonesia', TEFLIN Journal, 33(1), 47-74.

García, O. (2009). Education, multilingualism and translanguaging in the 21st century. In: Ajit Mohanty, A., Panda, M., Phillipson, R., and Skutnabb-Kangas, T. (eds). Multilingual

Education for Social Justice: Globalising the local. New Delhi: Orient Blackswan, pp. 128-145.

Garcia, O., Johnson, S. I., and Seltzer, K. (2016) The translanguaging classroom: Leveraging student bilingualism for learning. Philadelphia, PA: Caslon Publishing.

Garcia, M.H., Schleppegrell, M.J., Sobh, H., and Monte-Sano, C. (2023) ‘The translanguaging school’, Phi Delta Kappan, 105(2), 8-12

Gilquin, G. (2022) ‘Translanguaging and data-driven learning: How corpora can help leverage learners’ multilingual repertoires’, Teaching English as a Second or Foreign Language Journal–TESL-EJ.

Khairunnisa and Lukmana, I. (2020) ‘Teachers’ Attitudes towards Translanguaging in Indonesian EFL Classrooms’, Jurnal Penelitian Pendidikan 20 (2), pp. 254-266.

Krismanti, N. (2022) ‘Students’ Perspectives on Translanguaging in Classroom Context’, TEFLA Journal (Teaching English As Foreign Language and Applied Linguistic Journal), 4(1), pp. 8-15. https://doi.org/10.35747/tefla.v4i1.284.

Li, W. (2018) ‘Translanguaging as a practical theory of language’, Applied Linguistics, 39(1), 9-30.

Raja, F., Suparno, S., and Ngadiso, N. (2022) ‘Teachers’ attitude towards translanguaging practice and its implication in Indonesian EFL classroom’, Indonesian Journal of Applied Linguistics, 11(3), 567-576. https://doi.org/10.17509/ijal.v11i3.38371

Rasman. (2018) ‘To translanguage or not to translanguage? The multilingual practice in an Indonesian EFL classroom’, Indonesian Journal of Applied Linguistics 7 (3), 687-694.

Hickey, T. (2001) ‘Mixing beginners and native speakers in minority language immersion: who is immersing whom?’, The Canadian Modern Language Review, 57, 443-474.

- 11:25 **Paula Wood Borque** : Development of materials for the English as Foreign Language classroom: Compilation and analysis of a multimodal corpus

Abstract

It is well attested that human communication is essentially multimodal, as we orchestrate different modes to make meaning, especially nowadays with the rapid advancement in communication technologies and the media (Donaghy 2019). Because of this reason, there is a need to bring multimodal resources in the foreign language classroom to develop students' multimodal communicative competence (Royce 2006).

Audiovisual materials such as films and series are multimodal resources, as they combine different modes (linguistic, aural, visual, gestural, spatial) to communicate. Due to their multimodal nature, and because of their many benefits as a source of simulated naturalistic speech (Bednarek 2018), they can constitute a valid, rich source on which to develop materials for foreign language teaching. It will be argued that there is a need to compile multimodal corpora (Allwood 2009) of audiovisual materials such as films and series to be used as an informed source for the creation of materials to improve students' multimodal communicative competence in the English as a Foreign Language (EFL) Secondary Education classroom.

In this paper I will present a multimodal corpus of scenes from films and series in English, named CAMELLS (Corpus of Audiovisual Materials for English Language Learning in Secondary), and its process of compilation. The results from the multimodal analysis of 5 scenes from a film in CAMELLS will be presented in order to illustrate how the corpus can be used for the design of materials to be brought to the EFL classroom. In particular, I will examine the scenes' salient verbal and non-verbal features in order to convey certain meanings and meet specific communicative purposes and how to draw students' attention to them. It will be argued that the results obtained should inform the scenes' exploitation in the classroom to develop students' multimodal communicative competence.

References:

Allwood, J. (2009). Multimodal corpora. In Lüdeling, Anke & Kytö, Merja (eds.). "Corpus Linguistics. An International Handbook." Mouton de Gruyter: Berlin. (207-225).

Bednarek, M. (2018). "Language and Television Series: A Linguistic Approach to TV Dialogue." Cambridge: Cambridge University Press.

Donaghy, K. (2019). Using film to teach languages in a world of screens. In Herrero, Carmen & Isabelle Vanderschelden (eds.). "Using film and media in the language classroom: Reflections on research-led teaching." Multilingual Matters. (3-19)

Royce, Terry D. (2006). Multimodal communicative competence in Second Language contexts. In Royce, Terry D. & Wendy Bowcher (eds.). "New directions in the analysis of multimodal discourse." New York: Routledge. 361-390."

1045 - Talks 1, Syndicate 2

Chair: Chris Fitzgerald

- 10:45 **Valentin Werner, Robert Fuchs, Anna Rosen, Lea Bracke, Bethany Stoddard** : Introducing the Corpus of Young German Learner English

Abstract

Although Learner Corpus Research (LCR) has contributed significantly to a better understanding of Second Language Acquisition (SLA) processes in general, its full potential for the analysis of interlanguage (Selinker 1992) is yet to be realized. This is due to several major challenges identified in the literature (e.g. Tracy-Ventura et al. 2021). These include, among others, (i) an underrepresentation of beginner and lower-intermediate learners, (ii) an underrepresentation of spoken material and truly bi-modal data (i.e. data in different modes produced by the same learners), (iii) a lack of or unsystematic elicitation of metadata, (iv) a lack of longitudinal or at least quasi-longitudinal perspectives, and (v) a neglect of task effects.

The project presented in this poster will address these challenges by compiling and analyzing a corpus of Young German Learner English (YGLE). English is taught as a compulsory (mostly, first) foreign language to the vast majority of secondary schools students in Germany. Despite the important role that the teaching of English plays in the German education system, relatively little representative information is available on the overall learning outcomes, common learner errors and learning trajectories. This research gap can be addressed by LCR, which has the potential to provide representative and reliable information on the described target group of learners (Mukherjee 2008).

The YGLE corpus project thus aims to complement the extensive body of work on highly advanced L1 German EFL learners (e.g. Fuchs et al. 2016; Romer et al. 2020), based on various corpora of learners at the university level, by creating a database on the production of beginning to intermediate L1 German EFL learners in institutional contexts.

To this end, data are being collected from around 700 participants at secondary schools (learners aged 10-18 years) across the German three-tier school system. To represent a wide range of communicative contexts, the task types administered include both established (timed argumentative essay, picture description) and innovative task types (group discussion, elicitation of digital communication) with varying degrees of planning and interactivity. In addition, an extensive set of metadata is collected, based on a modified version of the questionnaire and procedure proposed by Moller (2017) and in line with the core L2 metadata scheme (Frey et al. 2023). This metadata comprises established and validated test batteries assessing information on socioeconomic and educational status, linguistic background, language use across different social contexts (including exposure to English outside of school), motivation (standardized tests FLM 3-6 R, FLM 7-13; Lohbeck & Petermann 2019; Petermann & Winkel 2015), as well as general and verbal cognitive abilities (standardized test AID-G; Kubinger & Hagenmuller 2019).

After transcription and annotation with a focus on items relevant for the complexity-

accuracy-fluency (CAF) triad, interactions between the CAF components as well as the influence of contextual and learner variables will be assessed using mixed-effects regression modeling. YGLE will eventually be made available to the LCR community , allowing (i) the exploration of areas beyond CAF (e.g. phonology, learner pragmatics, etc.) and potentially (ii) comparison with data from beginner and intermediate learners of English worldwide.

References

Frey, J. C., Konig, A., Stemle, E. W., & Paquot, M. (2023). A core metadata schema for L2 data. Paper presented at EuroSLA 32, Conference of the European Second Language Association. August 2023, Brimingham.

Fuchs, R., Gotz, S. & Werner, V. (2016). The present perfect in learner Englishes: A corpus-based case study on L1 German intermediate and advanced speech and writing. In V. Werner, E. Seoane & C. Suarez-Gomez (Eds.), Re-Assessing the Present Perfect (pp. 297-337). Mouton de Gruyter.

Kubinger, K. & Hagenmuller, B. (2019). Gruppentest zur Erfassung der Intelligenz auf Basis des AID. Hogrefe.

Lohbeck, A. & Petermann, F. (2019). Fragebogen zur Leistungsmotivation fur Schulerinnen und Schuler der 3. bis 6. Klasse - Revision. Hogrefe.

Moller, V. (2017). Language Acquisition in CLIL and Non-CLIL Settings: Learner Corpus and Experimental Evidence on Passive Constructions. Benjamins.

Petermann, F. & Winkel, S. (2015). Fragebogen zur Leistungsmotivation fur Schuler der 7. bis 13. Klasse. Pearson Harcourt.

Romer, U., Salicky, S. C. & Ellis, N. C. (2020). Verb-argument constructions in advanced L2 English learner production: Insights from corpora and verbal fluency tasks. Corpus Linguistics and Linguistic Theory, 16(2), 303-331.

Selinker, L. (1992). Rediscovering Interlanguage. Longman.

Tracy-Ventura, N., Paquot, M. & Myles, F. (2021). The future of corpora in SLA. In N. Tracy-Ventura & M. Paquot (Eds.), The Routledge Handbook of Second Language Acquisition and Corpora (pp. 409-424). Routledge.

- 11:05 **Rachele De Felice, Kate Warwick** : A new dataset for email research and professional communication: Fauci2020

Abstract

As is well known, real-world datasets of email communication are very rare, due to commercial and personal sensitivities. Research has been - and continues to be - carried out on emails from the Enron corporation and the Hillary Clinton administration, but there is a need for more recent and varied data against which to verify those findings. The appearance of any new potential public email datasets is therefore a significant event in linguistics research. One likely new source are the over 2700 messages sent to and from Dr Anthony Fauci between January and June 2020 (the first six months of the COVID-19 pandemic), obtained by the Washington Post, Buzzfeed News, and CNN through Freedom of Information Act (FOIA) requests. As one of the leading medical figures at the height of the pandemic, Fauci exchanged emails with a wide range of individuals including government officials, U.S. and international medical experts, the media, well-known individuals, and even the general public. This variety of interlocutors and the recency of the emails makes the dataset a promising source of material for research on email communication, professional language, and politeness. This presentation explores the viability of this corpus, considering issues including formatting, redactions, and range of correspondents. Related work by Benson et al. (2022) provides a useful starting point, although one more focused on the needs of computer scientists rather than of linguists. An exploratory study of basic politeness norms in this dataset then looks at canonical direct and indirect requests (I want, I need, can/could you, I was wondering, etc.), where findings can be easily compared to existing research. Initial results show a surprising low frequency of indirect requests, and relatively more imperatives, often modulated by please. The presentation will address these patterns in greater detail, with reference to hierarchy and social distance as well.

Austin R. Benson, Nate Veldt and David F. Gleich (2022) ‘fauci-email: a json digest of Anthony Fauci’s released emails’. Zenodo. doi: 10.5281/zenodo.5828209.

- 11:25 **Mateus De Souza, Michael McCarthy** : Exploring response tokens across proficiency levels: a spoken learner corpus study

Abstract

Until recently, learner corpus research has almost exclusively used data that is calibrated to the year-of-study of the learner, which can prove rather arbitrary when comparing across cohorts of first language groupings. The Common European Framework of Reference for Languages (CEFR) is becoming the international standard for language competence, and more corpora of learner data are being built so as to align with it, thus moving away from the year-of-study model. Although much has been done on written corpora calibrated to the CEFR, research on calibrated spoken corpora is in its nascence. This paper presents a brief description of the design of a new spoken learner corpus, with more than 100 transcribed recordings benchmarked to the CEFR. Additionally, it focuses on an investigation that describes learner listener behaviour through response tokens that learners use to show good listenership across proficiency levels. According to O'Keeffe et al. (2007, p. 142), response tokens are "vocal, verbal and non-verbal non-floor-holding devices that a listener may use to respond to the floor-holding message in a conversation". The methodology takes form-to-function and function-to-function approaches with quantitative and qualitative methods. Good listenership is an under researched area in spoken learner language, with implications for pedagogy that must be addressed for students to successfully communicate in the English variety they choose to learn (McCarthy 2002; O'Keeffe et al. 2007).

1045 - Talks 1, Syndicate 3

Chair: Pascual Pérez-Paredes

- 10:45 **Christoph Rühlemann** : Does gesture expressivity contribute to emotional resonance in storytelling interaction?

Abstract

Storytelling is driven by emotion. Its key function is a meeting of hearts: a resonance in the recipient(s) of the storyteller's emotion towards the story events [1]. How emotions are expressed gesturally is still seriously underresearched. This paper focuses on the role of gestures in emotion expression and emotion resonance in storytelling. The data come from the Freiburg Multimodal Interaction Corpus (FreMIC), which features not only CA transcriptions of video-recorded talk-in-interaction but also Electrodermal Activity (EDA) data on storytellers and story recipients [2]. Specifically, the paper asks three questions: Does storytellers' gesture expressivity increase from story onset to climax offset (RQ #1)? Does gesture expressivity predict specific EDA responses in story participants (RQ #2)? How important is the contribution of gesture expressivity to emotional resonance compared to the contribution of other predictors of resonance (RQ #3)?

The analyses are based on 44 stories (collected in 9 recordings, total run time 7.55 hrs, with 949 gestures and 13 distinct participants) and annotated for variables that may potentially impact emotion arousal. These include (i) Protagonist (whether the story's protagonist is the storyteller vs. a non-present third person), (ii) Recency (whether the story events occurred far in the past vs. they are occurring at or close to storytelling time), (iii) Group_composition (whether groups were all-female, all-male, or mixed), and (iv) Group_size (whether the storytelling setting was dyadic or triadic). Further, gestures were examined for whether they did or did not co-occur with a quote (variable G_quote). The gestures were further coded on a micro-analytic level for gesture phases [3] as well as for seven gesture-dynamic parameters: (i) Size (SO), (ii) Force (FO), (iii) Character view-point (CV) [4], (iv) Silence during gesture (SL), (v) Presence of hold phase (HO), (vi) Co- articulation with other bodily organs (MA) and (vii) Nucleus duration (ND). The annotations were implemented using a binary scale and aggregated in the Gesture Expressivity Index (GEI). The Index computes for each gesture an average value across all true/false ratings; the Index values are stored in the variable G_expressivity, one of the key variables in the models. Interrater agreement for the coding of the GEI parameters was tested on c. 24% of all 1,004 gestures and ranged between 79% for Force (FO) and 94% for Character viewpoint (CV).

To account for response latency, EDA responses were examined during the duration of the gesture as well as 1.5 sec post-gesture; further, they were classified as specific (i.e., as indexing a stimulus-related emotional response) if they were larger than 0.05 microSiemens. Finally, resonating gesture were identified, i.e., gestures exhibiting concurrent specific EDA responses by two or more participants, resulting in a binary variable EDA_G_resonance, the dependent variable in the Random Forest model. The first model, which addresses RQ #1, was a mixed-effects model with a relative positional measure G_position_rel for each gesture in each story (independent variable) and G_expressivity (dependent variable). The model suggested that storytellers' gestures become more expressive from story onset to climax offset.

To adress RQ #2, a second linear mixed-effects regression model was constructed, with EDA_specific_response_binary as the dependent and G_expressivity as the independent variable,. This model suggested that increased gesture expressivity increases the probability of specific EDA responses.

To address RQ #3 a Random Forest (ntree=1,500, mtry=3) was constructed for emotional resonance (EDA_G_resonance) as outcome variable and the seven GEI parameters as well as six more variables as predictors (G_quote, Protagonist, Group_compose, Group_size, Role (storyteller or story recipient), and Recency). The RF exhibited a very good fit: according to a one-tailed exact binomial test, the model was significantly better than chance/baseline ($p < .001$), the (traditional) R2 was 0.8595833, and McFadden's R2 scored an excellent 0.3689019. All but one predictor (Role) were found to impact EDA_G_resonance. Analysis of variable importance showed Group_composition to be the most impactful predictor, followed by Recency, Group_size, ND (nucleus duration), Protagonist, FO (gesture force), SZ (gesture size), G_quote, HO (hold phase), CV (character viewpoint), and MA (multiple articulators). Inspection of ICE plots clearly indicated combined effects of individual GEI parameters and other factors, including Group_size (the probabilities that gesture force (FO), size (SZ) and, respectively, nucleus duration (ND) impacts EDA_G_resonance were higher in triads) and Group_compose (the probabilities that these parameters impact

EDA_G_resonance were much higher for all-men groups than for all-female and mixed groups). Fig. 1 shows an ICE plot depicting the effect on emotional resonance of gesture force (FO) interacting with group composition (Group_compose). Methodologically, this study opens up new avenues of multimodal corpus linguistic research by examining the interplay of emotion-related metrics and gesture at micro-analytic levels and using advanced machine-learning methods to deal with the inherent collinearity of multimodal variables. More good is expected to come from this fruitful combination of qualitative and quantitative research.

References

1. Stivers, T. (2008). Stance, Alignment, and Affiliation during Storytelling: When Nodding Is a Token of Affiliation. Res. Lang. Soc. Interact. 41,31-57.
2. Rühlemann, C. and A. Ptak. (2023. Reaching below the tip of the iceberg: A guide to the Freiburg Multimodal Interaction Corpus (FreMIC). Open Linguistics: https://doi.org/ 10.1515/opli-2022-0245
3. Kendon, A. (2004). Gesture: Visible Action as Utterance. Cambridge, MA: Cambridge University Press.
4. Beattie, G. 2016. Rethinking body language: How hand movements reveal hidden thoughts. London/New York: Routledge

- 11:05 **Liina Repo** : Investigating Document Internal Variation: Modeling Historical Registers and Assessing the Impact of Text Segments on Register Classification

Abstract

Complexity, text variety, and lack of register information hinder the utility of historical language databases. Registers, situationally defined varieties with specific purposes, are important predictors of linguistic variation (Biber 2012). While registers are often examined at document level, recent studies have shown that longer documents exhibit features from different registers in different sections due to shifts in e.g. audience or purpose (Worsham and Kalita, 2018; Egbert and Gracheva 2022). We study this document internal register variation in eighteenth-century English texts by exploring the impact of different text segments on classification and inspecting the linguistic characteristics associated with them.

First, we inspect how different text parts (e.g., beginnings vs. endings) affect register classification using a BERT-based text classifier. BERT utilizes deep learning to comprehend language context, aiding in the analysis of linguistic characteristics of text segments. We fine-tune the model with register-annotated Corpus of Founding Era American English (COFEA). For testing, we use a hand-annotated section of Eighteenth Century Collections Online (ECCO). Second, to inspect and compare the linguistic characteristics and key features

across different text parts and registers, we employ the Stable Attribution Class Explanation (SACX; Rönnqvist et al. 2022) method. With SACX, we can compare key features across different registers using keyword lists derived from input attribution (Integrated Gradients) from the BERT model.

Our findings suggest varying degrees of influence among different text segments on register classification. Echoing past findings (Laippala et al. 2023), particularly text beginnings seem to produce more reliable classification results. Furthermore, certain features exhibit connections with specific parts of documents, akin to genre markers (Biber and Conrad 2019), while others demonstrate a more pervasive influence across the document. Understanding the pervasiveness of these features and their stability across texts facilitates more robust classification results and a deeper understanding of linguistic variation in historical documents.

- 11:25 **Lilia Shevyrdyaeva** : Academic tribes and discursive identity: metadiscourse patterns in research articles in closely related life sciences

Abstract

Disciplinary academic writing as a form of knowledge construction undergoes continuous change over time reflecting processes both inside and outside of academia. Each disciplinary community establishes and shares genre conventions and pragmatic strategies reflecting, to a certain degree, the scientific research it conducts. Modern day's pressure to publish increases the value of metadiscourse markers as effective tools of making a paper accepted by the disciplinary community. Metadiscourse markers contribute to building a convincing argument by structuring a text, projecting the author's standpoint, engaging the audience, establishing credibility, etc (Hyland, 2005).

Previous studies have observed variation in patterns of metadiscourse markers both between and within disciplines (Gillaerts&Van de Velde, 2010; McGrath&Kuteeva, 2012, Cao&Hu, 2014, Hyland&Jiang, 2018). Drawing on Hyland's framework (Hyland, 2018), this paper examines how academic authors with different disciplinary expertise use metadiscourse markers in the introduction and discussion sections of their research writing to mark the epistemic stance and establish a relationship with their audiences. This paper compares three closely related disciplines representative of the genre conventions, narrative tradition and language use in the life sciences - Ecology, Genetics and Immunology - to describe the variation of metadiscourse patterns. To this end, three sub-corpora were compiled of research articles from top-ranking disciplinary journals - 85 papers each (160,000-180,000 words) - published in 2019-2021 and authored by L1 English speakers. Both interactive and interactional metadiscourse markers were analyzed.

This comparative corpus-based investigation describes the frequency and distribution of metadiscourse markers across the sections of research papers and identifies specific patterns characteristic of each sub-genre. Quantitative and qualitative analyses reveal interdisciplinary variation and similarities between three academic discursive traditions. The most informative interactional markers exhibiting distinct differences turned out to be self-mention and, predictably, hedges and boosters, whereas for interactive markers interesting correlations were observed for code glosses, evidentials and transition markers, particularly, in introductions.

Cao, F., & Hu, G. (2014). Interactive metadiscourse in research articles: A comparative study of paradigmatic and disciplinary influences. Journal of Pragmatics, 66, 15-31. https://doi.org/10.1016/j.pragma.2014.02.007

Gillaerts, P., & Van de Velde, F. (2010). Interactional metadiscourse in research article abstracts. Journal of English for Academic Purposes, 9(2), 128-139. https://doi.org/10.1016/j.jeap.2010.02.004

Hyland, K. (2005). Stance and engagement: a model of interaction in academic discourse. Discourse Studies, 7(2), 173-192. https://doi.org/10.1177/1461445605050365

Hyland, K. (2018). Metadiscourse: Exploring Interaction in Writing (Bloomsbury Classics in Linguistics). Bloomsbury Academic. McGrath, L., & Kuteeva, M. (2012). Stance and engagement in pure mathematics research articles: Linking discourse features to disciplinary practices. English for Specific Purposes, 31(3), 161-173. https://doi.org/10.1016/j.esp.2011.11.002

Hyland, K., & Jiang, F. K. (2018). "In this paper we suggest": Changing patterns of disciplinary metadiscourse. English for Specific Purposes, 51, 18-30. https://doi.org/10.1016/j.esp.2018.02.001

1145 - Break, Foyer

1200 - Talks 2, Plenary room

Chair: Elaine Riordan

- 12:00 **Carolina Amador-Moreno, Ana María Terrazas-Calero** : Framing lithium to attract stakeholders: using CL to investigate language and Ecology

Abstract

The emerging field of Ecolinguistics, which explores the role of language in the interactions between humans, other species, and the environment (Fill & Penz 2017), has benefitted from the application of corpus-analytical techniques (Poole 2022). A number of recent studies dealing with various environmentally-related topics have started to combine the use of Corpus Linguistics techniques with Discourse Analysis approaches in order to look at what Alexander (2009: 24) calls 'discourse engineering'. Particular attention has been paid to the analysis of corporate discourse and the linguistic practice known as 'greenwashing'.

This paper aims to contribute to the field of Ecolinguistics by focusing on the discourse surrounding the mining of lithium, which is currently seen as an alternative to contemporary, environmentally-damaging energies.

In 2020, the European Commission updated their Critical Raw Materials list to include lithium as one of the key materials the EU needs to obtain in order to bring security and sustainability to Europe, encouraging member states to legislate in favor of mining and treating lithium within the EU. In the context of Spain, Australian mining company, Infinity Lithium, has identified the Valdeflores-San José mine as an unexploited lithium mine. Located near the UNESCO World Heritage city of Cáceres, the mine, which was originally designed as open-pit, has been strongly opposed by locals and activist who see their socio-ecological landscape threatened.

Using corpus linguistics and corpus stylistics techniques, this paper explores the linguistic strategies Infinity Lithium utilizes in a corpus of English-medium magazine articles to address the locals' concerns and garner their support. The use of key words Europe, EU, and European will be stylistically analyzed using Stibbe's (2015) frame and metaphor identification method to identify how the company addresses stakeholders and linguistically projects their development (and their corporate image) as pioneering, employment-producing, and environmentally-friendly.

Key words: Ecolinguistics; Corpus Stylistics; Corpus Linguistics; Frames and Metaphors; Lithium Mining. Bibliography

Alexander, R. J. (2009), Framing Discourse on the Environment: A Critical Discourse Approach. New York and London: Routledge.

Fill, Alwin F. and Hermine Penz (eds.). 2017. The Routledge Handbook of Ecolinguistics. Abingdon: Routledge.

Poole, Robert. 2022. Corpus-Assisted Ecolinguistics. New York: Bloombury.

Stibbe, Arran. 2015/2020. Ecolinguistics Language, Ecology and the Stories We Live By. Oxford: Routledge.

- 12:20 **Robbie Love, Nele Põldvere** : Variation across spoken genres: comparing the Spoken British National Corpus 2014 and the London-Lund Corpus 2

Abstract

We present a case study which brings together two contemporary corpora of spoken British English: the Spoken British National Corpus 2014 (Spoken BNC2014; Love et al., 2017) and the London-Lund Corpus 2 (LLC-2; Põldvere et al., 2021). Both corpora comprise transcribed spoken discourse produced by L1 speakers of British English. The Spoken BNC2014 transcripts (totalling 11.5 million words) are derived from recordings gathered from 2012 to 2016, and the corpus exclusively samples the broad genre of casual conversation. The LLC-2 comprises 500,000 words from a variety of discourse contexts, and its recordings were gathered from 2014 to 2019. As synchronically overlapping samples of a national variety, the corpora are potentially complementary: while the Spoken BNC2014 has the advantage of size (relative to the LLC-2), it is assumed to be relatively homogeneous in terms of genre; on the other hand, the LLC-2 is considerably smaller but captures a much greater diversity of speech contexts.

Our case study evaluates the potential for using the corpora to supplement each other to gain a more comprehensive picture of contemporary spoken British English lexis than possible in isolation. Using multidimensional analysis (Biber, 1988; Nini, 2019), we plot the Spoken BNC2014 and LLC-2 texts against Dimension 1 (Involved vs. Informational Discourse) in order to evaluate the genre coverage of the corpora according to formality. We then identify a sub-sample of the most prototypically 'conversational' texts (those with high Dimension 1 scores) and compare how the features of these texts vary between the corpora, considering the potential influence of speaker gender. We explore two factors of potential variation – genre and speaker gender – in order to (a) explore the extent to which social variation is influenced by register, and vice-versa, and (b) methodologically evaluate the complementarity of the Spoken BNC2014 and LLC-2.

References

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

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

Nini, A. (2019). The Multi-Dimensional Analysis Tagger. In Berber Sardinha, T. & Veirano Pinto M. (eds), Multi-Dimensional Analysis: Research Methods and Current Issues, 67-94. Bloomsbury Academic.

Põldvere, N., Johansson, V., & Paradis, C. (2021). On the London–Lund Corpus 2: Design, challenges and innovations. English Language and Linguistics, 25(3), 459-483. https://doi.org/10.1017/S1360674321000186

- 12:40 **Pascual Pérez-Paredes** : The power of stories: multilingual corpora and critical discourse analysis

Abstract

Stories have been widely used in narrative research. De Fina & Tseng (2017) have highlighted the relevance of stories in identity construction by transnational migrants and how they reveal group formation and agency as well as institutions and power asymmetries. Multimodal digital stories are personally narrated multimedia fragments that allow migrants to "negotiate self-representations as central or agentive vis- à-vis lived experiences" (Alexandra, 2008: 110–111).

In the wake of the Brexit referendum, EU citizens have become an important area of analysis for CDA and CADS. EU citizens exemplify how anti-migrant, populist discourses can impact citizens' rights and national and international migration policies (Hidalgo-Tenorio, Benítez-Castro & De Cesare, 2019). Research has focused on the impact of Brexit and anti-EU discourse on migrants (Guma & Dafydd Jones, 2019) or the framing of Brexit actors in the media (Parnell, 2023). However, there is a lack of CADS research that examines the voices of EU citizens exercising their freedom of movement in different countries of the EU (Authors, 2024). Compiling a corpus for such analyses is the aim of our research.

This study discusses the data collection methods and the multilingual nature of the FOMATPLAY corpus. Following Ambrosini, Cinalli & Jacobson (2020), we sought to create in the interviews the space for an examination of rights, practices, solidarities and discourses of EU citizens across national and transnational borders. Our corpus contains 100 video-recorded testimonies totalling 60 hours of recorded material and around 400,000 words in French, Italian and Spanish. The corpus has been annotated using a multilevel taxonomy of categories that explore migration, citizenship, identity and bordering topics. Issues of comparability are discussed, including the use of L2 languages, the selection of the interviewees, and tertium comparationis in contrastive analysis. In this presentation I examine a subset of the participants in the corpus: Romanian citizens living and working in Spain. We will look at identity construction and self-representation (Paraschivescu, 2022) of various aspects of the process of separation from their country and their relocation to new countries in the EU. This presentation seeks to demonstrate how multilingual research can impact traditional monolingual approaches to the construction and analysis of corpora.

References

Alexandra, D. (2008). Digital storytelling as transformative practice: Critical analysis and creative expression in the representation of migration in Ireland. Journal of Media Practice 9(2): 101–112.

Ambrosini, M., Cinalli, M., & Jacobson, D. (2020). Research on Migration, Borders and Citizenship: The Way Ahead. Migration, Borders and Citizenship: Between Policy and Public Spheres, 295-305. Authors (2024)

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

De Fina, A., & Tseng, A. (2017). Narrative in the Study of Migrants. In Canagarajah, S. (Ed.) The Routledge handbook of migration and language. Routledge, pp. 381-396.

Guma, T., & Dafydd Jones, R. (2019). “Where are we going to go now?” European Union migrants’ experiences of hostility, anxiety, and (non-) belonging during Brexit. Population, Space and Place, 25(1), e2198.

Hidalgo-Tenorio, E., Benítez-Castro, M. Á., & De Cesare, F. (Eds.). (2019). Populist discourse: Critical approaches to contemporary politics. Routledge.

Hunston, S. (2022). Corpora in applied linguistics. Second edition. Cambridge University Press.

Parnell, T. (2023). The representation of migrant identities in UK Government documents about Brexit: A corpus-assisted analysis. Journal of Language and Politics, 22(1), 46-65.

Paraschivescu, C. (2022). Experiencing whiteness: Intra-EU migration of Romanians to Paris and London. In New Trends in Intra-European Union Mobilities. Routledge, pp. 153-171.

Taylor, C. (2014). Investigating the representation of migrants in the UK and Italian press: A cross-linguistic corpus-assisted discourse analysis. International journal of corpus linguistics, 19(3), 368-400.

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

1200 - Talks 2, Syndicate 1

Chair: Gabrielle Gaudeau

- 12:00 **Catherine Wong** : Sherlock Holmes in the Computer: Quantitative Stylistic Methods

Abstract

This paper showcases the examples of using computer-assisted methods in performing pragmatic analysis of the literary subgenre of crime and detective fiction in determination of the discourse pattern and event structure of suspense. The body of work of this study is based on the prototype of this literary effect: Sir Conan Doyle's The Adventures of Sherlock Holmes (1892), and it aims at identifying and evaluating the pragmatic evidence by applying Levinson's Pragmatic Meanings (2000) in analysing words/phrases and turn-taking patterns used by different characters (the detective vs his foil characters) in describing topics - it extracts, compares and differentiates the lexical patterns used for presenting the clues in the narrative, the red herrings educed by the foil characters and the evidence observed and deduced by the detective. The discussion carries on to the application of Brewer and Lichtensten's Structural-Affect Theory in Stories (1982) in analysing the event organisation and sequence of the aforementioned clues, red herrings and evidence between the initial event (the crime) and the outcome event (the truth). The ultimate goal of this paper is to derive a pattern of which suspense is created through controlling the pace of revealing the outcome, and in other words, delaying its readers' cognitive processing in realisation of the 'truth'.

- 12:20 **Virginia Mattioli** : Crossing Translation Universals research and Language Variety studies: explicitation of the optional conjunction "that" in British and American translated novels

Abstract

This study aims to cross Translation Studies research on Translation Universals (TU) and Contrastive Linguistics studies on Language Variety by examining the use of the optative conjunction "that" in a set of novels translated into British and American English, respectively. According to the hypothesis of explicitation, translations tend to be more explicit than original texts from any perspective (Baker, 1993), including syntax. A number of Translation Studies scholars analysed and compared comparable corpora representing original and translated texts in order to corroborate or refute such hypothesis focusing on an ample gamut of syntactic features, including prepositions, conjunctions and optional linguistic elements (Olohan and Baker, 2000; Olohan, 2001 and 2002). However, the combination of TU and language variety is still mostly unexplored. This study aims to fill such gap by comparing the use of the optative conjunction "that" in British and American English translations, trying to answer to the following research question: has

linguistic variety any impact on the tendency to explicitation? With this goal, a set of four comparable corpora representing original and translated novels in British and American English, respectively, is compiled and examined. The adopted corpus-based methodology includes three steps: firstly, previous literature is reviewed to determine in which cases the subordinate conjunction "that" can be omitted; then, the relevant cases are identified in each corpus through specific searches in the concordance lists with the help of semantic and POS tags and wildcards; finally, the results are compared across the corpora from a quantitative and qualitative perspective. The results corroborate the TU hypothesis and suggest that tendency to explicitation varies according to the English variety used. Actually, British English authors and translators seem to be more likely to omit the optional "that" than American ones. Such results contribute to both Translation Studies and Contrastive Linguistics suggesting the need to consider the linguistic variation in the TU examination.

References:

Baker, Mona (1993), "Corpus Linguistics and Translation Studies – Implications and Applications", Text and Technology: In Honour of John Sinclair, eds. Mona Baker; Gill Francis; Elena Tognini-Bonelli. Amsterdam y Philadelphia, John Benjamins: 233-50.

Olohan, Maeve (2001), "Spelling out the optionals in translation: a corpus study", UCREL technical papers, 13: 423-32.

Olohan, Maeve (2002), "Leave it out! Using a comparable corpus to investigate aspects of explicitation in translation", Cadernos de Tradução, 1/9: 153-69.

Olohan, Maeve; Baker, Mona (2000), "Reporting that in Translated English: Evidence for Subconscious Processes of Explicitation?", Across Languages and Cultures 1/2: 141-58.

- 12:40 **Adina Ioana Vladu, Claudio Rodríguez Fer** : Unraveling the Linguistic Tapestry: A Computational Exploration of José Ángel Valente's Poetry

Abstract

This study aims to delve into the rich and intricate poetic universe of José Ángel Valente, a seminal figure in contemporary Spanish poetry, through the lens of linguistic analysis. Using computational and corpus linguistics and natural language processing (NLP) techniques, our research examines Valente's work to illuminate the poet's stylistic traits, thematic areas, and lexical patterns across his poetic work in Spanish.

Our choice is motivated, on the one hand, by Valente's significance in the Spanish literature of the 20th century through his rich, varied, and coherent artistic production, and, on the other hand, by the absence of similar interdisciplinary studies: the poet's work has been studied extensively from different literary approaches, but never up to this point from a linguistic perspective.

Our primary research questions include: What are the stylistic traits of Valente's poetry and how does the poet's stylistic expression evolve across his literary career? What are the predominant themes and how are they expressed in Valente's poetry? How do specific lexical patterns (e.g., word frequency, collocation patterns) contribute to the thematic and stylistic fabric of Valente's work?

In our exploration, we utilize a suite of computational linguistics and NLP tools, endeavoring to offer fresh perspectives on the chosen poetic corpus. This research represents an initial foray into understanding the complex interplay between language technology and literary analysis applied to the poet's works, hoping to inspire further interdisciplinary studies in the realm of Valente's poetry and beyond.

1200 - Talks 2, Syndicate 2

Chair: Yuan Gao

- 12:00 **Sijie Mou** : Rhotacisation in the Beijing Dialect: A study of Erhua

Abstract

The Beijing dialect has changed dramatically since the 1980s (Lin and Shen, 1995; Zhang, 2023) because of, for example, rapid socio-economic development in China (Zhang, 2001) and consequent population mobility in the metropolis (Hou, 1998; Zhou, 2002). Within this changing dialect, the erhua feature – rhotacisation – plays a significant role in the language variation. In Chinese dialects, erhua is a phonological feature which occurs in syllable-final position. Erhua can change the words grammatically (changing verb to noun), semantically (giving diminutive meanings), pragmatically (expressing friendliness), and stylistically (showing informality) in the contexts. Moreover, erhua usage is potentially relevant sociolinguistically.

While erhua is widely used, there is no clear phonological definition of the term. For example, terms such as ‘rhotacisation’, ‘r-colouring’, and ‘retroflexion’ are used synonymously. Neither do we have a straightforward picture of how erhua use relates to social categories. Traditionally, sociolinguistic studies of erhua were generally based on old corpora and collected by reading word lists. This research is grounded in a new corpus and hopefully will shed light on the definition of erhua and its use in different age groups, genders and districts of Beijing.

To address the definitional and sociolinguistic gaps with regard to erhua, this research aims (I) to clarify the form(s) of erhua based on phonetic analyses via Praat, (II) investigate whether and to what extent gender, area, and social network link to erhua, and (III) explore if erhua indexes specific social groups in contemporary Beijing. Accordingly, an updated spoken corpus has been built via Elan – a dataset of 24 speakers assembled by judgment sampling, balancing gender, age, area, and socio-economic status; flexible semi-structured sociolinguistic interviews for inducing a less formal discourse; Social Network Scales adjusted according to working status; image-based word lists of possible erhua occurrences. The extensive corpus provides a suitable dataset for sociolinguistic analyses.

References

HOU, J. 1998. Phonetic File of Beijing Dialect, Shanghai, Shanghai Educational Publishing House.

LIN, T. & SHEN, J. 1995. Phonetic Differences of Erhua Rhyme in Beijing Dialect. Studies of the Chinese Language, 170-179.

ZHANG, Q. 2001. Changing Economy, Changing Markets - A Sociolinguistic Study of Chinese Yuppies in Beijing. PhD, Stanford University.

ZHANG, W. 2023. A Sociolinguistic Study on the Rhotic Accent of Beijing Dialect. MA, Beijing Language and Culture University.

ZHOU, Y. 2002. Research on Modern Beijing dialect, Beijing, Beijing Normal University Publishing House.

- 12:20 **Shangran Jin, Gwen Bouvier, Zhao Li** : Self-help and cybersuperstition on China's social media Xiao Hongshu: Moralizing and objectifying luck

Abstract

Self-help type of debates on social media seek to assist users in developing more psychological resilience, self-awareness, and mental flexibility so that they can better manage life's stresses and obstacles (Nehring, 2020). This is a generally under-researched area in the Global South, including China (Nehring & Kerrigan, 2020). Nevertheless, the Chinese saying "attracting good luck" has recently evolved into a popular catchphrase among self-help influencers and their followers on the social media platform Xiao Hongshu - the equivalent of Instagram in China. The movement encourages users to post highly symbolic images such as lucky dogs, koi carp and lotus flowers (all markers of good luck in China). Much like the chain letters of yore (Seljamaa; 2008; Voolaid, 2013), these posts have the intention of generating good luck among the participating users. In this presentation, we look at one such an example of a chain started by an influencer. The corpus-based linguistic analysis draws out some patterns of the posts, which is followed up by a multimodal critical discourse analysis to dig deeper into the kinds of worldviews that underlie this chain behaviour. The study finds that good luck is abstracted, simplified and reduced in complexity, connecting the self-help aspect of these chains to people working on themselves as projects, in a Neoliberal and individualistic sense. The influencers, on the other hand, benefit from monetizing the frenzy of 'attracting good luck'. The study concludes that, within the influencer/follower relationship, the rising middle classes in China are encouraged to put up with hardship in their lives, and concentrate on self-improvement in order to receive good luck. Here, good luck becomes a task that requires effort and the popular adage "god helps those who help themselves" is pushed by the influencers profiting financially off all this online engagement. However, this Neoliberal form of pursuing luck conceals the responsibilities of social institutions that should also care for us and obscures different opportunities and abilities among people. On the other hand, it advances the harshness of competition in China whilst giving a contemporary digital shape to the illusion of "equality for all".

1200 - Talks 2, Syndicate 3

Chair: Andrew Caines (pre-recorded talks)

- 12:00 **Saqib Aziz** : Use of Lexical bundles in academic writing in English by expert writers, L1 English students, and L2 English students of Applied Linguistics

Abstract

This study compared the use of lexical bundles in academic writing in Applied Linguistics across three corpora: expert writers, native students and non-native students. The expert corpus consisted of articles published in Applied Linguistics journals; the native student corpus consisted of MA dissertations of native English students who did an Applied Linguistics Masters in English universities. The non-native student data consisted of Applied Linguistics MPhil dissertations of Pakistani students who did their MPhil in Pakistani universities. The size of the three corpora were as follows: native student corpus:312981, non-native student corpus:502945, expert writers' corpus:505958. The highly frequent bundles used in the three corpora were categorized into structural and functional categories (Hyland, 2008). These bundles were analyzed quantitatively as well as qualitatively. Thefindings revealed that the expert writers were different from native and non-native students in their use of structural and functional bundles. The expert writers used more Phrasal bundles and more bundles for organizing the text than the two student groups. The expert writers also showed better control of bundles for hedging. The students, on the other hand, used more bundles for describing research. Occasionally, they used vague and informal bundles, especially for quantifying. The non-native Pakistani students used far more bundles for describing the procedures of research and used far more bundle tokens than the other two groups. This might be due to the larger size of their dissertations. Interestingly, most of the differences between expert and student writers in their use of bundles applied to both sets of students. This suggests that the main challenge for all students is learning the conventions of academic writing, rather than any problems linked to non-nativeness. Therefore, the appropriate use of bundles in academic writing might need to be taught more explicitly to both native and non-native students.

- 12:20 **Ilia Afanasev, Olga Lyashevskaya** : Combining string similarity measures and frequency-based metrics: a new approach to measuring language distance between Slavic lects

Abstract

String similarity measures and corpus-based approaches are among the latest innovations in computational phylogenetic linguistics (Jaeger, 2019). Each of them is efficient for a specific material: lexicostatistic lists (Holman et al., 2008) or large raw (Gamallo et al., 2017) and small preprocessed (Hua, 2022) corpora respectively. However, neither can deal with small raw corpora (less than 10,000 tokens) alone.

We propose a method that combines string similarity measures and frequency-based metrics for measuring language distance between the lects presented by small raw corpora. The main comparison object is character N-grams (Zelenkov & Segalovich, 2007; Kosmajac & Keselj, 2020), as larger units, such as tokens, are too sparse for this corpora size. DistRank, a metric that compares the frequency of N-grams that coincide between two lects, is a primary tool. Non-coinciding N-grams go through a selection by similarity via string similarity measures, Levenshtein distance (Levenshtein, 1966) and Normalised Jaro-Winkler distance (Gueddah et al., 2015). The DistRanks between the most similar N-grams are weighted with the respective string similarity measure. The resulting lists undergo hybridisation in two ways: weighting of their mean values or calculating the overall mean value. The first way also employs the normalisation by Sørensen coefficient (Sørensen, 1948). The resulting clusterisation uses a UPGMA tree (Sokal & Michener, 1958).

The primary material of the research is historical East Slavic corpora (Smolensk, Polack, and Novgorod gramotas), all limited to approximately 1,000 tokens. Two other datasets assist in cross-verification of the findings: the modern East Slavic dataset of Belogornoje, Megra and Zialionka lects, as well as the modern South Slavic dataset of Croatian, Slovenian, and Slovak standards. After the experiments, the next step is to linguistically re-assess the metrics and outline the further research steps, including the implementation of probabilistic hierarchical clustering.

References

1. Gamallo, P., Campos, J., Alegria, I., 2017, “From language identification to language distance”, Physica A: Statistical Mechanics and its Applications, 484, 152–162.
2. Gueddah, H., Yousfi, A., Belkasmi, M., 2015, The filtered combination of the weighted edit distance and the Jaro-Winkler distance to improve spellchecking Arabic texts, IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA), 1-6.
3. Holman, E., Wichmann, S., Brown, C., Velupillai, V., Müller, A., Bakker, D., 2008, “Explorations in automated language classification”, Folia Linguistica, 42, 331–354.
4. Hua, X., 2022, “BayesVarbrul: a unified multidimensional analysis of language change in a speaker community”, Journal of Language Evolution, 7(1), 40–52.
5. Kosmajac D., Keselj V. Language Distance using Common N-Grams Approach // 2020 19th International Symposium INFOTEH-JAHORINA (INFOTEH). 2020.- C. 1 - 6. Levenshtein, V.I., 1965, “Binary Codes Capable of Correcting Deletions, Insertions, and Reversals”, Soviet Physics Doklady, 10, 707–710.

6. Sokal, R. R., Michener, C. D.. 1958, “A statistical method for evaluating systematic relationships”, University of Kansas Science Bulletin, 38, 1409–1438.
7. Sørensen, T., 1948, “A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons”, Kongelige Danske Videnskabernes Selskab, 5 (4), 1–34.
8. Zelenkov Ju.G, Segalovich I.V., 2007, “Sravnitel’nyj analiz metodov opredelenija nechetkih dublikatov dlja Web-dokumentov”, RCDL’2007, 1-9.

- 12:40 **Katherine Oliva Ortolani** : The Dimensions of Fashion, a Multidimensional Study

Abstract

Fashion influences global society in many different ways, however, the verbal language of fashion has received little attention. According to Moeran (2004), ‘could not exist without language.’ Most studies looking at the verbal language of fashion focus on topics such as terminology (Koester & Bryant, 1991), cross-language borrowing (Balteiro & Campos, 2012), and social semiotics (Barthes, 1983, 2013). There is a lack of studies on fashion from a discourse perspective, with discourses being ‘ways of looking at the world, of constructing objects and concepts in certain ways, of representing reality’ (Baker & McEnery, 2015, p. 5). To fill this gap, this paper presents a study whose goal is to detect the major discourses in the domain of fashion. A register-diversified corpus of fashion-related texts was compiled, consisting of newspaper and magazine features, television programs, and social media posts. The corpus was analyzed using lexical multi-dimensional analysis (LMD; Berber Sardinha, 2019), which derives from the MD analysis framework (Biber, 1988). LMD Analysis enables the identification of the lexical parameters of variation across the texts, which are detected statistically using Factor Analysis. The factors are interpreted in terms of the underlying discourses (e.g. Berber Sardinha, 2020). The corpus was tagged using the TreeTagger. The counts were computed, normed, and entered in a factorial analysis in SAS. Each text in the corpus was scored on each of the dimensions, and the scores were compared across the different registers. The dimensions, which will be introduced and illustrated in the presentation, suggest fashion as discourse on health, medical knowledge, and disease prevention; fashion as an empirical science; object of dialogue; history, professional success, and conceptualization; and fashion as self-esteem enhancement.

1300 - Lunch, Foyer

1400 - Talks 3, Plenary room

Chair: Mateus De Souza

- 14:00 **Dawn Knight, Paul Rayson, Mo El-Haj, Nouran Khallaf, Ignatius Ezeani, Steve Morris** : FreeTxt: a corpus-based approach to bilingual free-text survey and questionnaire data analysis

Abstract

Qualitative free-text responses (e.g. from questionnaires and surveys) pose a challenge to many companies/institutions who often lack the expertise to analyse such data with ease. While a range of sophisticated tools for the analysis of text do exist, these are often expensive, difficult to use and/or inaccessible to non-expert users. Such tools also lack support for the analysis of Welsh and English text, which can be a particular challenge in the context of Wales, as survey respondents should always be given the opportunity to respond in English and/or Welsh. Co-designed/co-developed in collaboration with National Trust Wales, National Museum Wales, Cadw, WJEC, and National Centre for Learning Welsh, this presentation introduces a novel corpus-based toolkit and methodological approach for the systematic analysis and visualization of free-text data in response to these two key gaps in knowledge. During the presentation we will:

- provide a detailed reflection of the process of co-design/co-construction of FreeTxt,
- give a live demonstration of all functionalities of FreeTxt (including the meaning analysis, sentiment chart, summarisation, word cloud, word tree and meaning use and relationships tools),
- discuss the wider impact/potential of FreeTxt and its use in a range of public and private sector organisations, and
- outline some of the potential extensions of FreeTxt and insights into other meaningful applications of corpus-based methods and utilities.

This presentation will demonstrate how, by working in partnership, software engineers, natural language processing (NLP) experts and corpus linguists can effectively collaborate to answer real world problems. FreeTxt is now available to use by anyone in any sector in Wales and beyond, with the potential to support the analysis of both small-scale and more extensive datasets bilingually. For more information, visit www.freetxt.app~

- 14:20 **Christopher Fitzgerald, Justin McNamara, Anne O'Keeffe, Dawn Knight, Geraldine Mark, Sandrine Peraldi, Tania Fahey Palma, Fiona Farr, Ben Cowan, Svenja Adolphs** : Interactional Variation Online: Building and Analysing a Corpus of Virtual Workplace Meetings

Abstract

The COVID-19 pandemic heralded an unprecedented shift towards a new type of work environment online. Virtual meetings quickly became a somewhat standard medium through which colleagues interacted. This paper describes how recordings of such meetings were curated to form the Interactional Variation Online (IVO) multi-modal corpus and how this corpus was analysed to study a number of features of virtual meetings. Insights into the corpus construction process from data capturing to transcription and annotation, including the use automatic transcription software, will be presented. We will also showcase some of the findings and outputs which have direct relevance to workplace communication.

The utility and challenges associated with designing, curating and using virtual meetings as corpus data are outlined in this talk. The advantages include the capacity to easily record video and audio capturing multiple participants, while the challenges include limited visibility of participants and reliance on participant hardware and third-party software. In addition, one of the challenges with working with multi-modal data is in determining how to approach analysis. This paper describes how the IVO corpus was constructed to facilitate various 'ways in' to multi-modal data that have been successful in analysing multiple modes in the IVO corpus (which are documented in the domain specific exemplars available on the project website: https://ivohub.com/resources/). To achieve this, we present how this rich dataset was used to investigate backchanneling behaviour (both verbal and non-verbal), how emblematic gestures are used in this environment and how patterns of silence provide a blueprint for how meetings are structured. In addition, we summarise some of the practical insights gleaned and discuss their relevance in the current workplace. Overall, we aim to illustrate the benefits of building a multi-modal corpus in terms of enhancing our understanding of this pervasive medium.

- 14:40 **Justin McNamara, Anne O'Keeffe, Christopher Fitzgerald, Dawn Knight, Sandrine Peraldi, Geraldine Mark, Tania Fahey Palma, Fiona Farr, Benjamin Cowan, Svenja Adolphs** : Is that an old hand?' Corpus Insights into Managing Virtual Meetings

Abstract

Online communication via video platforms such as Zoom and Microsoft Teams has become a modern standard component of workplace interaction. The uptake of these platforms was accelerated by the necessity to work from home for many during the COVID pandemic leading to a widespread adoption of virtual meetings as a primary means of communicating with colleagues. The Interactional Variation Online (IVO) project assesses both verbal and non-verbal communication in a multi-modal corpus of recorded virtual meetings. From this, the project identified typical features of virtual meetings in comparison to face-to-face meetings to inform best practice in virtual meeting facilitation and communication.

This paper investigates the corpus to determine how chairs manage turns in meetings, when restricted by the physical constraints of the virtual environment. We outline how chairs navigate the nomination of participants, to progress through an agenda and how silence plays a role in opening the floor. Our analysis offers insights into the important role of nomination by the chair in turn management, access to the floor and ensuring inclusivity through nomination, as well as efficient meeting progression. The paper will also discuss the challenges posed by the virtual environment in corpus annotation and the analysis of nomination, especially in relation to the visibility of virtual gestures. Three overarching nomination patterns were identified in the analysis: a) direct nomination (i.e. using a participant's name or title; b) indirect nomination (i.e. naming the agenda item for which it is implicitly understood that a given participant is responsible for); and c) self-nomination (i.e. participants take a turn where the chair has not nominated or named an agenda item). The paper will showcase the value of using a multi-modal approach to corpus analysis in bringing impactful insights to workplace discourse.

1400 - Talks 3, Syndicate 1

Chair: Yuan Gao

- 14:00 **Jing Chen, Yun Liu** : Metadiscourse in Research Article Abstracts in Musicology: An English-Chinese Comparative Study

Abstract

Metadiscourse, a vital linguistic resource for conveying stance and guiding reader interaction, has gained significant attention in academic writing. While previous research has explored its cross-linguistic and cross-disciplinary aspects, the metadiscourse in the musicology remains underexplored. Grounded in Hyland's interpersonal model, this study investigates metadiscourse in research article (RA) abstracts from English-medium and Chinese-medium music journals.

Two corpora, each comprising 120 abstracts from high-prestige music journals between 2016 to 2021, form the data of this study. Complementing this quantitative analysis, an open questionnaire seeks subjective insights from Chinese music learners, shedding light on metadiscoursal differences between Chinese and English RA abstracts.

Key findings highlight the dominance of metadiscourse markers in the English sub-corpus, particularly in the interactive and interactional dimensions. Intriguingly, Anglo-American authors exhibit a preference for interactional metadiscourse, while their Chinese counterparts lean towards interactive elements. Functional analysis elucidates distinct purposes behind the employment of frame markers, code glosses, and transition markers in both linguistic groups. Notably, both groups share a penchant for writer-oriented hedges to express opinions tentatively, emphasizing emphatics over amplifying adverbs. Complementary questionnaire results underscore a potential deficiency in academic writing training among Chinese music learners, potentially influencing the lower prevalence of metadiscourse in Chinese abstracts.

This study not only contributes to the growing body of metadiscourse research but also holds pedagogical implications for the enhancement of academic writing instruction in the field of musicology within the Chinese context.

- 14:20 **Ziwei Guo, Yu Chen** : A neural network model for constructional priming (resonance) in Mandarin and American English interaction

Abstract

Naturalistic communication is intrinsically grounded in resonance, which involves the activation of shared characteristics across turns . This phenomenon dynamically occurs when interlocutors creatively co-construct utterances that are formally and phonetically similar to those of a prior speaker (Du Bois, 2014; Tantucci, V., & Wang, A., 2021, 2024). The present study posits that the degree of similarity between turns can contribute to the machine learning prediction of linguistic and cross-cultural diversity. To test this hypothesis, the study analysed two balanced Callhome corpora, each containing 1,000 exchanges involving (dis)agreement in Mandarin Chinese and American English. The study found a correlation between the overt use of pragmatic markers and resonance, suggesting that resonance underlies dialogic engagement and cooperation among speakers. The researchers used a neural network model to demonstrate that resonance manifests differently across languages in terms of form and function. This study's applied results can contribute to a novel turn in AI research on conversational interfaces, as it reveals the fundamental role played cross-linguistically by resonance as a form of engagement in human-to-human interaction. It also highlights the importance of addressing this mechanism in machine-to-human communication.

References:

Du Bois, J. W. (2014). Towards a dialogic syntax. Cognitive Linguistics, 25(3), 359-410.

Tantucci, V., & Wang, A. (2021). Resonance and engagement through (dis-) agreement: Evidence of persistent constructional priming from Mandarin naturalistic interaction. Journal of Pragmatics, 175, 94-111.

Tantucci, V., & Wang, A. (2024). British Conversation is Changing: Resonance and Engagement in the BNC1994 and the BNC2014. Applied Linguistics,amae040.

- 14:40 **Hu Haiping** : A Corpus-Based Study on Translator's Style——Based on Translated Versions of Tao Te Ching by Arthur Waley and John Minford

Abstract

Tao Te Ching is one of the representative books of traditional Chinese culture, and according to statistics, Tao Te Ching is second only to the Bible in terms of the number of foreign translations published in the world. Based on the methodologies of corpus translation studies, this paper adopts quantitative and qualitative research methods to study the translator's style of sinologists Arthur Waley and John Minford through their translation of Tao Te Ching. The research questions of this paper mainly include the following four aspects: Firstly, What are the purposes motivating them to translate the Tao Te Ching into English? Secondly, What kind of readers are they targeting in their translation? Thirdly, What are the translation styles of them embedded in the texts? Fourthly, How did they design their paratexts?

Research results reveal that the styles of the two sinologists were different in both the pre-translation and while-translation phases. In the pre-translation stage, firstly, the common English cultural background and consciousness of the two sinologists prompted both of them to choose to stand on their cultural stance and defend their cultural values during translation. Secondly, they have different purposes for translating the Tao Te Ching. Waley translates the Tao Te Ching for academic purposes, placing Taoist thought in the context of the ancient Chinese philosophical thought system, while Minford hopes that the Tao Te Ching can inspire the readers' attitudes towards life. Thirdly, the different motives for translating the Tao Te Ching created different readers' orientations: Wiley mainly served anthropologists, while Minford was oriented to the general public.

At the translation stage, the translation styles of the 2 sinologists have their characteristics, and the statistics of the corpus tool WordSmith Tools 8.0 show that, in terms of average sentence length, Waley's version is longer, and in terms of vocabulary richness, the vocabulary used by Minford is richer. In terms of the treatment of paratexts, Waley favors academization, and Minford's paratexts are full of rich Chinese elements. All these features are closely related to their translation purposes.

1400 - Talks 3, Syndicate 2

Chair: Gabrielle Gaudeau

- 14:00 **Michelle Zeping Huang, Kacey Jianwen LIU** : Discourse of Menopause on Social Media across Cultures: A Corpus-Assisted Comparison between X and Weibo

Abstract

This research aims to examine how menopause is represented on social media across cultures. Specifically, the present study adopts a novel corpus-assisted approach combining with critical discourse analysis to compare social representations of menopause between two popular social media platforms: Twitter (now renamed X) and Weibo. X, predominantly using English, and Weibo, a Chinese-language platform widely used by Chinese speakers, serve as two distinct cultural contexts for study. Social media posts containing the keywords menopause and 更年期(geng nian qi; menopause in Chinese) published between 2018 and 2023 are retrieved from X and Weibo respectively. Traditional corpus linguistics methods including frequency and concordance analysis are initially applied to identify the top 50 frequent words and the top 100 collocates of menopause / 更年期(geng nian qi) in each corpus. A comparison analysis is then conducted to examine the similarities and differences in representations of menopause between the two corpora. To complement the corpus linguistics findings, we incorporate topic modelling and sentiment analysis to uncover major themes related to menopause and users' attitudes towards menopause. This study contributes to the existing literature on cross-cultural studies of menopause on social media. The findings shed light on the public's perceptions and sentiments towards menopause. It also provides insights to medical practitioners, healthcare communicators and researchers to further offer support and professional information to middle-aged women who are going through perimenopausal or menopausal periods.

- 14:20 **Wenxuan Ren** : An UFA-approach to diachronic investigation into the identity construction of vegetarian in China Daily

Abstract

This study aims to reveal the construction of the image of vegetarians and vegans in China Daily and the diachronic change in the usage of the words *vegetarian* and *vegan* by using the corpus-based Usage Fluctuation Analysis (UFA) (McEnery et al. 2019) method to examine collocates around the word *vegetarian* and *vegan* in a 535 thousand word corpus of newspaper articles in China Daily published between 2001 and 2023, expecting to outline the diachronic relation between the shift of collocates and the usages of the target words, thereby considered exploring how culture guides people to adopt healthy diets and how individual choices of diet shape health standards. The results are as follows: Firstly, two types of collocates of *vegetarian* and *vegan* can be distinguished: the consistent collocate and the transient collocates. While the consistent collocate only has the word *diet*, various transient collocates are extracted and are divided into the following categories (in order of frequency): food, lifestyle, religion, and health. The transient collocates of *vegetarian* and *vegan* have some similarities and differences. The similarities lie in the strong connection with food and the gradual shifts from a dietary requirement closely related to religious beliefs to a healthy choice of diet. The difference includes the time of the first occurrence of the word, the degree and frequency of the usage fluctuation, and the categories of collocates. Secondly, based on collocates, in China Daily, vegetarians and vegans are constructed as individuals who choose a particular diet with a certain type of lifestyle. Before 2018, there was a stronger tendency to associate them with religious beliefs in Buddhism or Taoism. Finally, the dominant diachronic trend in the construction of vegetarians and vegans is mainly influenced by major worldwide incidents in health and hygiene, religious events, and the international vegetarian and vegan movement.

Reference:

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

- 14:40 **Katrien Deroey, Jane Helen Johnson** : Importance marking in EMI and L1 lectures

Abstract

In this talk we compare lecture discourse by English Medium Instruction (EMI) and English L1 lecturers with regard to how they signpost important points. Identifying important points is arguably a key aspect of effective lecture delivery and comprehension (Jung, 2003; Kiewra 2002; Revell and Wainwright 2009). An engineering lecture corpus was examined to identify 'importance markers', i.e. lexicogrammatical devices that overtly mark the importance of verbal or visual points (Deroey, 2015).

The corpus of 46 lectures (ca. 365,000 words) is composed of four subcorpora: L1 UK and New Zealand lectures from the Engineering Lecture Corpus (ELC) (www.coventry.ac.uk/elc) and Malaysian and Italian EMI lectures from the ELC and Bologna University Corpus of Engineering lectures (EmiBO) (Picciuolo & Johnson, 2020). The transcripts were read to identify and tag importance markers and the tagged transcripts were uploaded into SketchEngine (http://www.sketchengine.eu/) for further analysis. The corpus yielded 376 importance markers, most of which could be classified into verb, noun and adjective patterns depending on their core lexeme (see Deroey & Taverniers, 2012). Assessment references (Johnson, 2024) and idioms formed additional categories.

Our analysis revealed three main findings. First, the overall picture emerging from the comparison is one of similarity. The EMI and L1 lecturers marked importance to the same extent (187 EMI; 189 L1), preferred verb and premodified noun markers (e.g. 'remember'; 'my main point is') and used a similar variety of patterns and lexis. Second, where substantial differences were found in the frequency of particular patterns or lexis, this was due to idiolect. Third, the EMI and L1 subcorpora showed considerable internal variation.

We will conclude by highlighting the following aspects of our findings: insights into EMI lecture discourse; implications for the construction of EMI corpora; and implications for lecture listening and lecturer training.

References

Deroey, K. L. B. (2015). Marking importance in lectures: interactive and textual orientation. Applied Linguistics, 36(1), 51-72. doi:10.1093/applin/amt029 (Advance online publication 2013)

Deroey, K. L. B., & Taverniers, M. (2012). 'Just remember this': Lexicogrammatical relevance markers in lectures. English for Specific Purposes, 31(4), 221-233.

Johnson, J. H. (2024). "This is for sure a question at your exams": Assessment references in EMI and L1 engineering lectures. ESP Today, 12(1), 155-177.

Jung, E. H. S. (2003). The role of discourse signaling cues in second language listening comprehension. The modern language journal, 87(4), 562-577.

Kiewra, K. A. (2002). How classroom teachers can help students learn and teach them how to learn. Theory into Practice, 41(2), 71-80.

Picciuolo, M., & Johnson, J. H. (2020). Contrasting EMI lecturers' perceptions with practices at the University of Bologna. In Miller, D.R. (Ed.), Quaderni del CeSLiC. Occasional papers. Bologna: Centro di Studi Linguistico-Culturali (CeSLiC), Università di Bologna. AlmaDL, p. 23. http://amsacta.unibo.it/6399/

Revell, A., & Wainwright, E. (2009). What makes lectures 'unmissable'? Insights into teaching excellence and active learning. Journal of Geography in Higher Education, 33(2), 209-223.

1400 - Talks 3, Syndicate 3

Chair: Diana Galvan-Sosa

- 14:00 **Cathryn Bennett** : Employing learner needs in corpus literacy teacher education (CLTE) programmes

Abstract

The call for corpus literacy training in teacher education has been recently renewed (Farr & Lenko-Szymanka, 2023), in efforts to boost its use in mainstream teaching practice. While several studies have concentrated on student-teachers becoming corpus literate in pre-service training programmes, fewer have focused on in-service teachers in the private ELT sector (Xodabande & Nazari, 2023). It has been argued that providing corpus literacy training to this highly motivated group of teachers in challenging work conditions can minimise lesson preparing time (Bennett, 2023) and increase job satisfaction (Bandura, 1977).

Previously, research has paired teachers' needs in the classroom with how these can be met with corpora (Römer, 2009; Callies, 2019), while others advocate for the learner to be more involved in the materials design process (Clarke, 1991). Thus, an in-service corpus literacy training framework was designed utilising learner needs (Bennett & Uí Dhonnchadha, 2023). Although corpus resource books were consulted, the aim of the training framework was to train teachers how to design materials based on their students' needs.

Over a two-year period, 19 EFL practitioners integrated corpus activities into their EFL classroom. Optional ready-made corpus materials were provided and based on participants' learners' needs. Learner needs were collected via reflective journals that were completed by participants in the training programme.

In terms of grammar, the most common learner needs included how to use conditionals and prepositions, while vocabulary needs involved English for Specific Purposes. Depending on their students' levels, teachers designed direct/indirect activities. The results of the study call for learners to be more involved in CLTE programmes.

References

Bandura, A. (1977) Social learning theory (Vol. 1). Prentice Hall.

Bennett, C. (2023) Investigating the experiences of in-service English language teachers in the use of language corpora for teaching purposes: An international action research study. [Unpublished doctoral thesis]. Trinity College Dublin.

Bennett, C., & Uí Dhonnchadha, E. (2023) Becoming Corpus Literate: An in-service EFL teacher education framework for integrating corpora into EFL teaching. Applied Corpus Linguistics. 3(1). https://doi.org/10.1016/j.acorp.2023.100048.

Callies, M. (2019) Integrating corpus literacy into language teacher education. Learner corpora and language teaching, J. Mukherjee and S. Götz, (Eds.) 245-263. John Benjamins.

Clarke, D.F. (1991) The Negotiated Syllabus: What is it and How is it Likely to Work? Applied Linguistics. 12(1). 13-28. https://doi.org/10.1093/applin/12.1.13.

Farr, F., & Lenko-Szymanska, A. (2023) Corpora in English Language Teacher Education: Research, Integration, and Resources. TESOL Quarterly. https://doi.org/10.1002/tesq.3281.

Römer, U. (2009) Corpus research and practice: What help do teachers need and what can we offer? In Corpora and Language Teaching. John Benjamins. 83-98.

Xodabande, I., and Nazari, M. (2022) Impacts of a Corpus Linguistics Course on in-Service EFL Teachers' Corpus Literacy. CALL-EJ. 23(1). 318-346.

- 14:20 **Sepideh Daghbandan** : Conversational Language Forms used by Learners of Persian: Insights from a Learner Corpus-based Study

Abstract

Conversational Persian is at its early stages of receiving attention in the field of Teaching Persian as a Second Language. However, research on the use of the Conversational Persian by learners of Persian remains scarce. Therefore, this study aims to explore the use of Conversational Persian by language learners using a learner corpus.

To this end, a spoken learner corpus, namely, the Learner of Persian Spoken Corpus (LoPSC) was compiled. LoPSC is the first spoken corpus collected from learners of Persian. Data from LoPSC consists of approximately 40,000 words of transcribed audio recordings from conversations between advanced learners of Persian. After the compilation of LoPSC, to gain a better understanding of the challenges that learners may encounter when using Conversational Persian, LoPSC was compared to a reference corpus, namely, the Conversational Persian Corpus. This corpus consists of 60,000 words of audio-transcribed recordings from conversations of Persian speakers living in Iran.

The results from the corpus-based analysis revealed that the most significant difference between the use of Conversational Persian by learners and first language speakers of Persian

was in their use of discourse markers. That is, the learners used significantly fewer discourse markers compared to their L1 speaker counterparts. The two groups of speakers also used different pragmatic functions for the same discourse markers.

This study has three main contributions. First, it provides empirical findings in a novel context, namely, the use of Conversational Persian by learners. Second, this study also provides further empirical evidence on how learners use discourse markers in Conversational Persian, especially in comparison with L1 speakers of Persian. Finally, as the first study to compile and analyse a spoken learner corpus in Persian, this study also provides insights into the challenges of compiling a learner corpus in this language, especially regarding the conversational register of Persian.

- 14:40 **HE Liang** : Researching Poverty-themed Children's Picture Books in the Contemporary United States: An Approach of Multimodal Discourse Analysis

Abstract

Poverty is a global issue. Even the most economically developed countries in the world today have not been able to completely eradicate poverty. Due to differences in many aspects such as education, economy, politics, and history, the education of poverty values differs among countries. A proper understanding and correct method to combat poverty is an important part of cultivating children to face up to social problems and establish values. Picture books, as the "first book in life" and an important medium for transmitting values, play an indispensable role in children's poverty education. Also, as a multimodal discourse, these books have the function of representing and constructing social reality, and promoting social change. However, the existing picture book research lacks attention to the theme of poverty, with a few studies mainly focusing on qualitative analysis, and research on poverty-themed picture books for specific countries is deficient.

In view of this, this study focuses on 15 poverty-themed American children's picture books from 2000 to 2020. Corpora including images and words are built, and computer-aided text mining technology is used. Qualitative and quantitative research is conducted based on a five-level multimodal discourse framework, with the facilitation of Antconc3.5.9, UAM Corpus Tool, and UAM Image Tool. This study constructs a multimodal discourse analysis system for children's picture books, which is composed of subsystems such as "the multimodal resources system", "the semiotics interaction and cooperation mechanism system", and "the socio-cultural context and value interpretation system". It involves levels of semiotic modes, structures, semantics, context, and value.This study aims to analyze the theme of poverty in picture books, and helps children to understand poverty and establish correct poverty values.

1500 - Break, Foyer

1530 - Panel Discussion, Plenary room

Topic: Trust the Text? Generative AI and Corpus Linguistics

Discussants: Michael McCarthy, Paula Buttery, Dawn Knight, Michael Handford

Chair: Andrew Caines

1630 - Poster Session & Drinks Reception

Chairs: Diana Galvan-Sosa, Yuan Gao, Gabrielle Gaudeau

- **Yueming Du, Shue Sum Leung** : Do features of lexical richness distinguish and predict Chinese-as-a-Second-Language writing quality?

Abstract

Introduction: This study investigates the impact of lexical richness on Chinese-as-a-Second-Language (CSL) writing quality, addressing gaps in current research. Previous studies on lexical richness in CSL writing were limited by small-scale corpora and manual extraction of features, which may not adequately represent lexical richness.

Aims: The study aims to explore the relationship between CSL writing quality and lexical richness, covering four dimensions of lexical richness based on Read (2000): diversity, sophistication, density, and errors. It seeks to identify which features vary across different levels of writing and determine the best predictors of L2 writing quality.

Methodology: The study utilized the HSK Dynamic Composition Corpus, comprising graded, error-coded CSL writing scripts from candidates in 20 countries/regions. A total of 3000 texts (1000 each from high, mid, and low levels based on specific mark boundaries) were analyzed. Using LTP 4 (Che et al., 2021) for preprocessing and a Python program, 64 features corresponding to the four aspects of lexical richness were extracted.

Results: One-way ANOVAs indicated significant differences in 61 measures across different script levels. Notably, high-level scripts excelled in 11 indices of lexical diversity (e.g., word types), mid-level scripts differed in 18 indices of lexical sophistication (e.g., low-frequency words), four measures of lexical density significantly differentiated mid-level scripts, and five lexical error measures varied between high and mid-level scripts. A stepwise regression revealed that 8 variables, including word type count and high-frequency vocabulary portion, explained 44.14% of the variance in writing marks.

- **Luís Martínez-Kleiser Magaña** : Online dissemination of scientific content for educational purposes: recontextualising texts for teenage audiences

Abstract

Democratization of knowledge seeks to render complex concepts available to a broader audience, empowering individuals to create a personal vision of the world and to construct informed opinions. Due to the spread of digital media, there are young adults who enjoy swift access to scientific information which requires recontextualization for effective communication. The existence of blogs or websites disseminating scientific knowledge adapted and made comprehensible for teenagers is a fact. It is the goal of this study to curate a corpus of digital scientific publications designed for young learners, probing into the verbal and non-verbal strategies employed.

To dissect the features and how scientific-specialized language is recast and constructed for adolescents, a small-scale corpus of 10 texts published between 2020 and 2023 on the site Science Journal for Teens has been selected. These samples target 12 to 18-year-old English native speakers in Middle School and High School and endeavour to make science research-based concepts accessible to such ages. The topics span from pollution, possible uses of waste and recycling, climate change impact, environmental laws, ecological farming or global warming to mental and physical healthy habits.

The investigation encompasses discursive strategies such as paraphrasing, elaboration, and condensation of information. Given the prevalence of multimodal communication in the digital landscape, attention will be directed toward the analysis of the orchestration of modes, especially the verbal and visual ones, looking into layout, colour selection, and the integration of images and graphics. The ultimate aim is to identify patterns that facilitate the transformation of intricate findings intended for expert audiences into digestible concepts for young minds. Recontextualizing scientific texts for teenage audiences extends beyond enhancing comprehension; it involves the development of persuasive linguistic and multimodal strategies to captivate young readers on pertinent topics. Authors strive for this engagement while upholding knowledge legitimacy and credibility.

The diversity of modes employed to bring technical knowledge closer to young audiences and the way they impact understanding by such readers will be discussed together with the potential use of these texts in English as a Foreign Language Secondary Education setting, aiming to facilitate the dissemination of scientific content for educational purposes.

- **Nazym Shaikhina** : Shaping Perspectives: Unraveling Gender Narratives in Kazakhstani Education through the Lens of “Ozin-ozi tanu” Textbooks

Abstract

This study critically examines the portrayal of gender roles in the “Ozin-ozi tanu” textbooks used in Kazakhstan from 2010 to 2022. “Ozin-ozi tanu,” a course introduced by Sarah Nazarbayeva and taught in primary and secondary schools, aims to promote self-discovery and moral development through the teaching of universal human values. By analyzing four textbooks, this research identifies the frequency and context in which male and female characters are depicted, highlighting significant gender biases. Using corpus linguistic tools like AntConc, the study conducts both quantitative and qualitative analyses to uncover recurring themes and narratives. The quantitative analysis focuses on the frequency and distribution of male and female characters, while the qualitative analysis explores the context and connotations associated with these characters by examining collocates. The analysis shows that women are consistently described using terms such as “family,” “nurturing,” and “child,” emphasizing their roles within domestic and familial contexts. Conversely, men are characterized with descriptors like “leader,” “warrior,” and “honour,” highlighting their association with leadership and military attributes. The study also reveals that women are notably underrepresented in mentions and references within the texts. Many female characters lack substantial presence or individual names, emphasizing their exclusion from significant roles and narratives compared to their male counterparts. This nuanced exploration through corpus linguistics uncovers how gender roles and stereotypes are subtly reinforced within educational materials. This research underscores the impact of educational materials on shaping gender perceptions and calls for a critical reevaluation of these portrayals to promote gender equity. The paper aligns with Strand 2 – Corpus Linguistics, Pragmatics, and Discourse at the conference, highlighting the role of corpus methods in critically examining educational texts.

- **Lingmin Huang, Yuanke Li** : Revisiting the relationships of n-gram measures to L2 writing proficiency: Comparisons between genres and connections to vocabulary levels

Abstract

In an emerging line of L2 writing research, some natural language processing tools (e.g., CollGram, TAALES) have been employed to automatically analyze which n-grams, namely contiguous multiword sequences, used by EFL learners in their compositions appear in large reference corpora such as the BNC or COCA and measure their frequency, association strength and range in these corpora. Some of these n-gram measures have been found to strongly correlate to L2 writing proficiency (e.g., Garner et al., 2019; Granger & Bestgen, 2014; Kim et al., 2018; Li & Fang, 2022; Zhang & Li, 2021). However, the extant studies have primarily focused on one genre, i.e. argumentative essays (e.g., Bestgen & Granger, 2014; Chen, 2019; Garner et al., 2018, 2019), leaving much room to investigate other important genres. Additionally, these studies have mainly revolved around word frequency (e.g., Bestgen & Granger, 2014; Garner et al., 2019; Granger & Bestgen, 2014) and "specialized" vocabulary (Gablasova et al., 2017, p. 10). Whether and how critical n-gram measures predictive of English writing proficiency have any connection to vocabulary levels in the English curriculum has not been investigated.

Against this background, we investigated (1) whether there were similar or varied n-gram measures predictive of holistic scores among two genres (argumentative essays and request letters) and (2) the relationship between specific multiword combinations connected to the n-gram measures predictive of writing scores and their vocabulary levels in the Chinese national English curriculum. To this end, 60 indices of bigrams and trigrams were employed to analyze 600 rated English essays and 600 rated English request letters composed by Chinese high school students. The results showed that three pairs of similar n-gram measures predicted the holistic scores among these two types of writings, but five indices were uniquely associated with letter scores. In both genres, higher mean MI associations in more proficient compositions were significantly correlated to more multiword combinations containing words from college and graduate levels in advanced writings than in intermediate writings.

References

Bestgen, Y., & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. Journal of Second Language Writing, 26, 28–41. https://doi.org/10.1016/j.jslw.2014.09.004

Chen, A.C. (2019). Assessing phraseological development in word sequences of variable lengths in second language texts using directional association measures. Language Learning, 69(2), 440–477. https://doi.org/10.1111/lang.12340.

Gablasova, D., Brezina, V., & McEnery, T. (2017). Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. Language Learning, 67(S1), 155–179. https://doi.org/10.1111/lang.12225.

Garner, J., Crossley, S., & Kyle, K. (2018). Beginning and intermediate L2 writer's use of N-grams: An association measures study. International Review of Applied Linguistics in Language Teaching, 58(1), 51–74. https://doi.org/10.1515/iral-2017-0089.

Garner, J., Crossley, S., & Kyle, K. (2019). N-gram measures and L2 writing proficiency. System, 80, 176–187. https://doi.org/10.1016/j.system.2018.12.001

Granger, S., & Bestgen, Y. (2014). The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. International Review of Applied Linguistics in Language Teaching, 52(3). https://doi.org/10.1515/iral-2014-0011

Kim, M., Crossley, S.A., & Kyle, K. (2018). Lexical sophistication as a multidimensional phenomenon: Relations to second language lexical proficiency, development, and writing quality. The Modern Language Journal, 102(1), 120–141. https://doi.org/10.1111/modl.12447.

Li, Y. K., & Fang, A. B. (2022). Synergistic effects of multiword sequences structure, function, frequency and association on raters' evaluations of essay quality. Frontiers in Psychology, 13, 1026658. https://doi.org/10.3389/fpsyg.2022.1026658

Zhang, X., & Li, W. (2021). Effects of n-grams on the rated L2 writing quality of expository essays: A conceptual replication and extension. System, 97, 102437. https://doi.org/10.1016/j.system.2020.102437

- **Yixing Liu** : China's Media Representations of Vegetarianism: A Corpus-Assisted Discourse Analysis

Abstract

This study conducts a comparative analysis of the reporting on vegetarianism in the Chinese and English editions of "China Daily" to explore differences between domestic and international communication. Given the increasing global concerns over health issues, such as rising obesity rates and environmental sustainability, dietary choices have become a focal point of interest, making this research particularly relevant. The study employs a combination of Corpus-Assisted Discourse Studies (CADS) and Critical Discourse Analysis (CDA) to systematically analyze linguistic patterns, collocations, and concordances in the portrayal of vegetarianism in both editions of "China Daily."

The aim is to uncover the diverse representations of vegetarianism in the media and understand the ideologies and power dynamics embedded in these portrayals. By comparing the domestic and international narratives, the study investigates the role of media in shaping public health awareness and dietary preferences, especially in the context of the global obesity issue.

We expect the findings to offer a comprehensive understanding of the portrayal of vegetarianism in the Chinese and English editions of a major Chinese media outlet, shedding light on the relationships between language, power, and ideology in shaping public perceptions. This study contributes to academic discourse and has implications for public health policies and media practices relating to health topics. The analytical approach, combining quantitative

and qualitative analyses, offers an in-depth understanding of media discourse with the potential to challenge and reshape societal norms and values regarding dietary choices.

- **Ridha Rashed Alanazi** : Exploring The Use of Formulaic Language in English Learning: A Corpus-Based Study of Arabic Students at a Saudi University

Abstract

This study investigates the use of specific types of formulaic language, namely phrasal verbs and collocations, by Arabic students learning English at a Saudi university. A corpus, created from these students' written texts, will be analyzed to understand their use of these language forms. The analysis will involve identifying frequently used words, phrasal verbs, and collocations, as well as examining patterns of words that frequently appear together.

The objective is to understand how the use of phrasal verbs and collocations changes, especially as students progress academically. It is hypothesized that more advanced students might exhibit a greater command of these language forms. In January 2023, a pilot study was conducted using established study instruments, including the Phrasal Vocabulary Size Test (PVST) by Martinez (2011), to measure students' knowledge of formulaic language. The pilot study provided some insights, particularly regarding the test administration. It was observed that students tended to score higher on the PVST when they had access to online resources. This raised concerns about the accuracy of the test results in reflecting the students' true understanding of formulaic language. Consequently, to ensure a more reliable assessment, future administrations of the PVST will be conducted under controlled conditions that restrict access to the Internet.

The research currently focuses on male students due to constraints in collecting data from female students at the university. This is a recognized limitation of the study. In its current phase of data collection for the main study, the research aims to uncover patterns in the acquisition and use of formulaic language, specifically phrasal verbs, and collocations. The findings are expected to provide valuable insights for improving teaching methods and curricula in English language education, particularly tailored to Arabic-speaking learners.

Keywords: Formulaic Language, Phrasal Verbs, Collocations, English Learning, Arabic Students, Language Education Research

- **Charles Lam** : Sub-disciplinary Variations in Biology Texts and Implications on Instructions

Abstract

Variations in organisation of academic writing are well documented across disciplines (Samraj 2002; McGrath & Kuteeva, 2012; Lu et al., 2021), or even among branches of the same discipline (Samraj, 2005; Ozturk, 2007, Kanoksilapatham 2015). To complement existing studies on the correlation between rhetorical moves and lexical bundles or syntactic complexity (Staples et al., 2013), the present study analysed 50 publicly available research articles from BioRxiv, evenly distributed in five categories: Animal Behaviour, Biochemistry, Biophysics, Ecology, and Physiology. Each sentence was human-annotated by steps in the CARS model (Swales, 1990).

The data show that Move1-Step3 “Reviewing previous studies” is the most frequent step (42% of all steps). Variations between the five sub-disciplines are also found using two metrics: co-occurence of steps and lengths of the steps. By plotting all the steps, this study found that Move1-Step3 “Reviewing previous studies” occurs typically near M1_S2 “Making topic generalizations” and M2_S2 “Indicating a gap”, with M2_S2 being the most common option among the four in Move2. The plot also reveals that M3_S1a “Outlining purposes” and M3_S1b “Announcing present research” may occasionally appear earlier, supplementing discovery in previous studies that M1_S3 is often interspersed with other steps. The second metric, length of steps, refers to how many sentences a rhetorical step may contiguously span over. Biochemistry articles dedicate more on M1_S3 (51.1%) than other disciplines (Animal Behaviour: 32.4%, Biophysics: 43.2%, Ecology: 41.6%, Physiology: 47.3%).

Using publicly available sources allows instructors to easily find realistic examples that are suitable for the teaching context. In addition to the data-driven learning element, the use of realistic data encourages students to learn from articles in their respective disciplines in terms of writing too.

References

Kanoksilapatham, B. (2015). Distinguishing textual features characterizing structural variation in research articles across three engineering sub-discipline corpora. English for Specific Purposes, 37, 74-86.

Lu, X., Casal, J. E., Liu, Y., Kisselev, O., & Yoon, J. (2021). The relationship between syntactic complexity and rhetorical move-steps in research article introductions: Variation among four social science and engineering disciplines. Journal of English for Academic Purposes, 52, 101006.

Ozturk, I. (2007). The textual organisation of research article introductions in applied linguistics: Variability within a single discipline. English for specific purposes, 26(1), 25-38.

McGrath, L., & Kuteeva, M. (2012). Stance and engagement in pure mathematics research articles: Linking discourse features to disciplinary practices. English for Specific Purposes, 31(3), 161-173.

Samraj, B. (2002). Introductions in research articles: Variations across disciplines. English for specific purposes, 21(1):1–17.

Samraj, B. (2005). An exploration of a genre set: Research article abstracts and introductions in two disciplines. English for specific purposes, 24(2):141–156.

Staples, S., Egbert, J., Biber, D., & McClair, A. (2013). Formulaic sequences and EAP writing development: Lexical bundles in the TOEFL iBT writing section. Journal of English for academic purposes, 12(3), 214-225.

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

- **Jing Huang, Tanjun Liu** : Digital mediation of romantic relationships in China: Scholarly educational videos' comments on Bilibili

Abstract

The rocketing divorce rate along with the plummeting marriage rate in mainland China in recent years evoked the public concern. Some academics keenly participated in the public discussions, created contents on popular social media and shared ideas with the public. Language, especially of experts, is instrumental in establishing categories of difference, relations of inequality, or social norms by which we feel obliged to live our lives (Thurlow, 2018), while power in discourse is likely to receive a boost in social media (KhosraviNik, 2014). This study focuses on one well-known sociologist's four-episode educational serials on the topic of "getting out of singlehood" from 2021 to 2024, on one of the most popular youth-oriented video-sharing platforms in China, Bilibili. The series had more than 16000 comments and in total more than 0.5 million tokens. This research investigates Shen's and the audience's construction of romantic relationships and their views on getting out of singlehood through analyzing collocations of high frequency words including love (爱情aiqing) and getting out of singlehood (脱单tuodan). The results show that the audience's agreement with Shen's dismissal of those who likened romantic relationships to an idealized "love model" and decided whether someone they met could be fit in the model. The audience also tended to prioritize the interests and needs of self over the other party in dating or developing relationships. These findings signal an individualistic trend and resonate with the surge of egoism (Yan, 2003) in relationship discourse in current China. This research contributes to the relationship of mediatisation and the transforming discourses about love, relationship and dating in China. It helps clarify the practices of co-constructing such discourses, regarding the impact of the academic influencers upon audiences and the agency of media participants.

- **Qianhui Sun** : Hands-off data-driven learning for Chinese students of English at the basic-independent levels

Abstract

Data-driven learning (DDL) has been demonstrated to be a helpful method in the teaching and learning of collocations and lexico-grammatical patterns (Lee et al., 2019; Muftah, 2023). Nevertheless, the available literature has indicated that learners with limited language proficiency may perceive DDL as daunting and overwhelming (Kennedy & Miceli, 2001, 2010; Yoon & Hirvela, 2004). In response to this challenge, the present ongoing research project focuses on the implementation of paper-based DDL, termed 'hands-off DDL,' as a way of simplifying learners' interaction with corpus data while still encouraging them to discover language patterns by themselves. More specifically, this research project examines the applicability of hands-off DDL for students of English at the basic and independent levels (i.e., at A2 and B1 CEFR levels) at a public university in China. Over a six-week instructional period, a general English course will integrate hands-off DDL activities designed by the researcher. The case study will evaluate students' attitudes towards hands-off DDL, the learning strategies they employ in class, and the effects of hands-off DDL in their learning autonomy and strategy use three months after the pedagogical intervention. Data will be collected through (i) entry questionnaires; (ii) semi-structured classroom observations for six weeks; (iii) semi-structured interviews with focal students during the intervention (iv) an exit questionnaire to be administered to all students; (v) a delayed questionnaire to be completed by all students three months after the intervention; and (vi) interviews with focal students three months after the intervention. The findings of the present research project are expected to inform the adoption of hands-off DDL in language classrooms in China for students with limited language proficiency, an area of research which has not received much attention to date in the academic literature on DDL.

- **Agustina Lestary** : A Corpus-Based Analysis of Teachers' Questions in an EFL (English as Foreign Language) Classroom in Higher Education Context in Indonesia

Abstract

There have been many studies conducted related to teachers' questions in Indonesian context, but they are mostly conducted in secondary level context. In addition, studies about teachers' questions focus the discussion only on certain type of questions (e.g display questions or referential questions). Taken into consideration that secondary level education and higher education settings are different in nature (such as the objectives of classroom teaching), this study would extend the existing literature on teachers' questions by focusing on the description of different types of questions used by teachers and their pedagogical roles in higher education context in Indonesia. Self-Evaluation of Teacher Talk (SETT) Framework by Steve Walsh is employed to categorise the classroom modes and the questions. The data is collected by audio-recording an EFL (English as Foreign Language) classroom in a university in Banjarmasin, Indonesia. NVivo is used to classify the classroom modes and to categorise the questions used by the teacher. Sketch Engine is used to identify the most frequent lexical bundling used by teacher in their questions. The findings of the data suggest that teacher' questions could help the students relate to the material presented in students' presentation by discussing their own personal experiences. Some lexical bundling such as 'what else' and 'any other opinion' could be used as cue for students to contribute to the discussions.

- **Tanjun Liu** : Investigating formulaic sequences in university students' disciplinary writing

Abstract

Formulaic sequences play a crucial role in language use, processing, and acquisition, constituting a vital component of language competence (Erman & Warren, 2000). Previous studies have shown that formulaic sequences serve as noticeable disciplinary markers (Hyland, 2012). However, proper use of formulaic sequences in academic writing, particularly disciplinary writing, continues to pose a challenge for L2 learners, even at relatively higher proficiency levels (e.g., Chen & Baker, 2010; Laufer & Waldman, 2011).

Therefore, this study aims to investigate the use of formulaic sequences in university-level EFL learners' disciplinary writing. A corpus consisting of students' disciplinary writing is built. English writing samples were collected from undergraduates majoring in four disciplinary areas, namely business, humanities and social science, (natural) science, and advanced technology, at an English-medium instruction university in China. Students contributed their writing for the module courses they wrote each academic year. Based on corpus analysis using Sketch Engine, the preliminary results indicate some similarities

in the use of certain formulaic sequences between two disciplinary areas, business and humanities and social science. More significantly, the results provide further evidence of the divergent use of formulaic sequences across disciplines in academic prose. Additionally, it is observed that L2 learners face challenges in using some formulaic sequences and demonstrate limited variety in their usage. Interestingly, there was some development over time in the use of formulaic sequences. These findings contribute to our understanding of improving L2 learners' use of academic formulaic sequences in different disciplines. The study also has pedagogical implications for both disciplinary module teachers and English language teachers, emphasising the potential benefits of collaboration in teaching.

Reference list:

Chen, Y. H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. Language Learning & Technology, 14(2), 30-49.

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. Text & Talk, 20(1), 29–62.

Hyland, K. (2012). Bundles in academic discourse. Annual review of applied linguistics, 32, 150–169.

Kilgarriff, A., Baisa, V., Busta, J., Jakubicek, M., Kovár, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. Lexicography, 1, 7–36.

Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: a corpus analysis of learners' English. Language Learning, 2(61), 647–672.

Day 2: Wednesday 17th July

0900 - Welcome, Registration, Coffee

In the foyer of the Crausaz Wordsworth Building, Robinson College, Cambridge.

0930 - Keynote 2, Plenary room

Dr Geraldine Mark, Cardiff University, U.K.

Bio: Geraldine Mark is a corpus linguist whose interests include language development, data-driven learning, and discourse analysis. Most recently she has been working on a multi-modal corpus project, IVO (www.ivohub.com), examining multi-modal interaction in the virtual workplace. She is a Visiting Lecturer at the University of Malta, and advises on the FoRCE project, building and analysing a corpus of Maltese English. She is co-author of *English Grammar Today* (2011, Cambridge University Press, with Ronald Carter, Michael McCarthy and Anne O'Keeffe) and co-principal researcher (with Anne O'Keeffe) of the *English Grammar Profile*, an online resource profiling L2 grammar development.

Title: **Shall I compare thee to the BNC?**

Abstract: Comparison is a cornerstone of corpus linguistics. Comparisons can be superficially simple, but a scratch below the surface typically reveals complexity, when variation between corpora and within corpora is not always obvious, and where "there is little understanding of what constitutes a small or large difference between corpora" (Gablasova et al. 2017). This talk is a personal reflection on some of the pitfalls and affordances encountered when comparing one corpus with another, particularly in relation to learner language. I draw on three case studies, firstly using the Cambridge Learner Corpus* to look at development in writing across the Common European Framework of Reference (CEFR) proficiency levels, secondly contrasting adverb use in spoken language from the LINDSEI and LOCNEC corpora[†], and finally exploring the design and development of a corpus of spoken and written Maltese English. I touch on measures of language development in terms of quality and time (Durrant et al. 2021), the use norms and attitudes to variation within learner corpus research, and consider implications for corpus design and representativeness.

References:

Durrant, P., Brenchley, M., & McCallum, L. (2021). *Understanding development and proficiency in writing: Quantitative corpus linguistic approaches.* Cambridge University Press.

Gablasova, D., Brezina, V., & McEnery, T. (2017). Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. *Language learning*, 67(S1), 155-179

* The Cambridge Learner Corpus (CLC) is a corpus of written exam data from Cambridge English exams. The data used for this study stands at 55.5 million words and comprises 266,600 exam documents of open-ended writing, spanning 143 different first languages backgrounds, six proficiency levels, from a 17-year period (1993–2012).

† LINDSEI (Louvain International Database of Spoken English Interlanguage); LOCNEC (Louvain Corpus of Native English Conversation)

Chair: Prof Paula Buttery

1030 - Break, Foyer

1045 - Talks 4, Plenary room

Chair: Andrew Caines

- 10:45 **Elizabeth Hanks, Tony McEnery, Jesse Egbert, Tove Larsson, Douglas Biber, Randi Reppen, Paul Baker, Raffaella Bottini, Vaclav Brezina, Gavin Brookes, Isobelle Clarke** : The Lancaster-Northern Arizona Corpus of Spoken American English (LANA-CASE): Design, compilation, and ongoing progress

Abstract

This presentation introduces the design and collection of a new corpus of American English conversation: the Lancaster-Northern Arizona Corpus of Spoken American English (LANA-CASE). In this ongoing project, we are compiling a publicly available, large-scale corpus of spoken American English which can serve as a counterpart to the BNC2014 (Brezina et al., 2021; Love et al., 2017).

This presentation describes the procedures utilized in its construction, focusing on innovative methods that may be adapted to future corpus creation projects. We discuss spoken corpus design, sampling, and data collection, along with a brief progress report of LANA-CASE.

We will cover the following:

1. Design: we design the corpus by following methods proposed in Egbert, Biber, and Gray (2022), including describing the domain, operationalizing the domain, planning the sample, and evaluating the sample;
2. Sampling: we adopt an iterative approach to sampling in order to consistently evaluate and improve the representativeness of the corpus. We utilize a sampling criteria framework following Love et al. (2017), in which several participant demographics guide sampling efforts (e.g., age, geographic region, gender) while more extensive metadata is also collected (e.g., interlocutor relationship, languages spoken);
3. Data collection: we leverage public participation in scientific research (Shirk et al., 2012) by collecting spoken data through online questionnaires that allow participants to contribute remotely. We utilize social media, public outreach efforts, and market research panels to recruit participants from diverse regions, race/ethnicities, ages, and settings (urban/suburban or rural);

4. Progress to date: we summarize the progress made, including a review of the data that has been collected (about 500 hours of conversation) and transcribed (about two million words), evaluating our data in relation to the design goals previously established.

- 11:05 **David Oakey, Franco Zappettini, Ziwei Guo** : Corpus creation with and without keywords – does achieving corpus “aboutness” introduce bias?

Abstract

In recent years a number of research articles have been published which describe and discuss the discourse around a particular topic, informed by specialised corpora of texts such as news stories and op-ed pieces about the topic in question. A common methodological procedure is the use by researchers of keywords to identify relevant articles to include in their corpora such as ‘Britain’, ‘Europe’ and ‘Brexit’, (Parnell 2023: 57); ‘coronavirus’, ‘COVID’, ‘at-risk’, ‘cases’ (Davies 2021: 588); and ‘Ukraine’ (Kryzhanivska 2022: 14).

Using pre-selected linguistic search terms to include or exclude texts from a corpus risks introducing bias into the data, however, so that results merely confirm and reinforce the researcher’s pre-conceived notion of which words are “about” a particular news story. Corpus pioneer John Sinclair warned researchers of “a vicious circle arising if they construct a corpus to reflect what they already know or can guess about its linguistic detail” and went on to stipulate a corpus design requirement that “the contents of a corpus should be selected without regard for the language they contain, but according to their communicative function in the community in which they arise” (Sinclair, 2005: 5). Furthermore, a keyword-based approach to identifying texts for a corpus can result in relevant linguistic features about the topic being missed altogether (Oakey et al 2022).

This paper discusses a study which aims to reveal to what extent linguistic information about a topic is likely to be missed by the study of a corpus collected using keywords. It will compare results from a keyword-based news article corpus “about” a particular news topic with results from a similar-sized corpus of news articles from the same time period for which keywords were not used. The comparison offers insights into the consequences of a particular choice of corpus construction method for a study’s results.

References

Davies, M. (2021). The Coronavirus Corpus: Design, construction, and use. International Journal of Corpus Linguistics, 26(4), 583-598. doi:https://doi.org/10.1075/ijcl.21044.dav

Kryzhanivska, A. (2022). The Beginnings: Russian-Ukrainian War in European, Ukrainian, and Russian Media in 2014 15. Balcania et Slavia, 2(1), 9-26.

Oakey, D. J., Jones, C., & O’Halloran, K. L. (2022). Phraseology and Imagery in UK Public Health Agency COVID-19 Tweets. In S. Tan & M. K. L. E (Eds.), Discourses, Modes, Media and Meaning in an Era of Pandemic: A Multimodal Discourse Analysis Approach (pp. 89-114). London: Routledge.

Parnell, T. (2023) Humiliating and dividing the nation in the British pro-Brexit press: a corpus-assisted analysis, Critical Discourse Studies, 20:1, 53-69, DOI: 10.1080/17405904.2021.1983446

Sinclair, J. M. (2005). Corpus and Text — Basic Principles. In M. Wynne (Ed.), Developing Linguistic Corpora: a Guide to Good Practice (pp. 1-16). Oxford: Oxbow Books.

- 11:25 **Laurence Anthony** : Enhancing Corpus Analysis through the Integration of Large Language Models (LLMs)

Abstract

In the realm of natural language processing research, Large Language Models (LLMs) have emerged as powerful tools that offer novel, surprisingly, profound, but also questionable insights into language usage across diverse domains and registers. These models are built using vast amounts of language data and are readily accessible through web interfaces and APIs. The challenge lies in understanding and evaluating LLM outputs, given their ‘black box’ design and their tendency to generate ‘hallucinations’ (inaccuracies in model output), especially when they lack representative data in a target domain. This paper addresses the challenges of using LLMs by seamlessly integrating them into a conventional corpus analysis toolkit and establishing a direct connection between the LLM and user-defined corpora. This integration enables users to perform targeted LLM-based queries about individual corpus files or the entire corpus, as well as prompting the LLM for insights on results generated by traditional corpus tools, such as KWIC concordancers and collocate tools. The integration of LLMs with corpus tools described in this paper also allows for strategic prompt engineering that significantly mitigates the risk of ‘hallucinations’. Moreover, the accuracy of LLM-derived insights can be easily validated using direct links to the original corpus data, thereby enhancing the credibility and utility of LLMs in corpus research.

1045 - Talks 4, Syndicate 1

Chair: Yuan Gao

- 10:45 **Christopher Fitzgerald, Ivor Timmis** : ‘Have your coffin ready near you for you will suffer death’ Comparing and contrasting Irish and English 19th century threatening letters

Abstract

This paper compares two corpora of anonymous threatening letters from the late eighteenth and early nineteenth centuries. One corpus is a collection of 520 letters from pre-famine Ireland and the second is a collection of 165 letters from 1760-1820 England. This paper draws on results generated from the application of corpus tools to highlight similarities and differences between the two corpora in terms of lexis and discourse and the strategies the writers adopted to meet the literacy challenge.

The two corpora share several characteristics. The main similarity is that their common aim is to threaten their recipient. In addition, both corpora contain numerous biblical invocations which help the writers to produce intelligible prose as well as allowing them to strike a tone of righteous indignation. Both sets of letters are comparable in their literary competence, creativity, figurative language and humour. While this might seem contrary to the destitute status of the authors, much of the authorship can be attributed to the impressive resourcefulness of the writers in repurposing language from familiar sources and support from more educated members of the community.

While the similarities between both corpora are remarkable, there are some features which distinguish the Irish letters e.g. a greater focus on nationalistic sentiment and the Catholic faith, and the use of features of Irish English. Attendees at this paper will encounter rich examples of orality in letters with the same aim, written circa the same era but in two different geographical and political contexts.

- 11:05 **Guyanne Wilson** : Agreement in African and Caribbean Englishes: Collective nouns and existential constructions

Abstract

Variation in agreement in English existential constructions and with collective nouns has been the focus of much research on Inner Circle Englishes, and particularly the standardised varieties spoken in the United Kingdom and the United States (e.g. Crawford 2005, Levin 2006). There have also been investigations of both of these phenomena in Asian and East Asian varieties of English (e.g. Collins 2012, Hundt 2009). However, with the exception of Jantos' (2009) work on agreement in Jamaican English more generally, neither of these syntactic forms has been explored in detail in African and Caribbean Englishes, though there is good reason to do so. Firstly, it would give a more comprehensive picture of agreement in English more generally. Furthermore, African and Caribbean Englishes are linked historically, yet are only rarely studied in relation to one another outside of Creole Linguistics. This work is one of the largest comparisons of African and Caribbean Englishes to date.

This paper discusses agreement in three African (Nigerian English, Ghanaian English, and Ugandan English) and three Caribbean varieties of English (Grenadian English, Trinidad and Tobagonian English, and Jamaican English). It starts by looking first at looking at agreement in non-existential, declarative sentences such as "The children are laughing", before turning to agreement in there+ BE existential constructions and with collective nouns such as "staff" and "team". The study sought to find out the form agreement in existential constructions and with collective nouns takes in African and Caribbean Englishes, the linguistic constraints acting on agreement, and the effect of register (spoken versus written) and text category on agreement.

For each variety, except Grenada, the component corpus of the International Corpus of English (ICE) was used in order to study the frequencies of agreement (e.g. There are five children) and non-agreement (e.g. There's five children) in existential constructions. It also explored the use of singular agreement (e.g. The government is) and plural agreement (e.g. The government are) with collective nouns. For Grenada, a smaller parallel corpus was compiled. Agreement was compared across registers (spoken versus written) as well as text types and text categories.

Overall, it was found that rates of non-agreement in existential constructions in African and Caribbean Englishes were markedly lower than those reported for other varieties of English in previous works. However, it was higher than non-agreement in non-existential constructions in the varieties under examination (i.e. constructions such as, "the children was laughing loudly"). With regard to collective nouns, African and Caribbean Englishes were found to display high levels of singular agreement, even with words that are plural only in other varieties, notably police. However, there do exist individual lexical differences. In this way, African and Caribbean Englishes appear to be more conservative both than other Outer Circle varieties of English and Inner Circle varieties of English in which there is greater variation.

References

Collins, Peter. 2012. Singular agreement in there-existentials: An intervarietal corpus-based study. English World-Wide 33(1). 53–68.

Crawford, William J. .2005. Verb Agreement and Disagreement A Corpus Investigation of Concord Variation in Existential There+ Be Constructions. Journal of English Linguistics 33.1, 35-61.

Hundt, Marianne. 2006. The committee has/have decided: On concord patterns with collective nouns in inner-and outer-circle varieties of English. Journal of English Linguistics 34(3). 206-232.

Jantos, Susanne. 2009. Agreement in Educated Jamaican English: A Corpus Investigation of ICE-Jamaica. Freiburg: University of Freiburg. (Doctoral Dissertation).

Levin, Magnus. 2006. Collective nouns and language change. English Language & Linguistics 10(2). 321-343.

- 11:25 **Valentin Werner, Hendrik Michael, Lea Bracke** : Authenticity and intimacy: A corpus study on live blogs about the US presidential debates

Abstract

The present contribution engages with the larger topic of discourse and politics through as-sessing live blogging (LB; Thurman & Walters, 2013) as a form of web-native (political) jour-nalism. Discourse practices in online reporting have been found to be characterized by hybridi-ty in terms of (i) sticking to strategic rituals of objectivity to create accountability on the one hand (e.g. Singer, 2005; Lasorsa et al., 2012) while (ii) also being marked by storytelling or "news as narrative" (Wahl-Jorgensen & Schmidt, 2019) to make the world more transparent, recognizable, and graspable. Relating to the latter in LB specifically, Tereszkiewicz (2014) discussed how it creates polyvocality by relying on amateur sources and featuring interactions between reporters and users. From a similar vantage point, Steensen (2016) described an "in-timization of journalism" by increasingly blurred boundaries between the personal and the professional perspective on social media.

The present study, which uses political LB about the 2020 presidential debates in the United States as a case in point, explores whether such LB represents polyvocal discourse involving the expansion of voices and perspectives, fostering authenticity through including everyday voices as sources and emphasizing transparency through the presence of a reliable narrator who regularly provides updates and fact-checks; and (ii) whether it simultaneously is a form of intimate discourse representing immediacy and emotionality, blending the professional and private roles of the communicator and the audience.

The study relies on a purpose-built corpus of LB coverage of the two televised US presidential debates (Donald Trump vs. Joe Biden). Data were collected from four popular media outlets (The Guardian, Daily Mirror; New York Times, Wall Street Journal). The overall corpus

size amounts to 61,490 tokens. To facilitate a discourse-oriented mixed-methods approach (Bedna-rek & Carr, 2021) a combined quantitative-qualitative analysis with AntConc and MAXQDA was conducted. For the operationalization of the hypotheses established categorizations from journalism studies (e.g. Donsbach & Klett, 1993; Bruns, 2018) to annotate sourcing practices, (multimodal) markers of transparency and authenticity, as well as (lack of) linguistic indicators of journalistic objectivity were applied.

As is visible both in sourcing and from a linguistic perspective, the results highlight the abovementioned practices of blending and the integration of new media practices that result in creating different modes of conveying information through storytelling and thus create jour-nalistic perspectives that are accountable, intimate and authentic.

References

Bednarek, M., & Carr, G. (2021). Computer-assisted digital text analysis for journalism and communications research: Introducing corpus linguistic techniques that do not require programming. Media International Australia, 181(1), 131–151.

Bruns, A. (2018). Gatewatching and news curation: Journalism, social media, and the public sphere. Lang.

Donsbach, W., & Klett, B. (1993). Subjective objectivity: How journalists in four countries define a key term of their profession. Gazette, 51, 53–83.

Lasorsa, D. L., Lewis, S. C. & Holton, A. E. (2012). Normalizing Twitter. Journalism Studies, 13(1), 19–36.

Singer, J. B. (2005). The political j-blogger. 'Normalizing' a new media form to fit old norms and practices. Journalism, 6(2), 173–198.

Steensen, S. (2016). The intimization of journalism. In T. Witschge, C. Anderson, D. Domin-go, & A. Hermida (Eds.). The Sage handbook of digital journalism. Sage, 113–127.

Tereskiewicz, A. (2014). "I'm not sure what that means yet, but we'll soon find out": The discourse of newspaper live blogs. Studia Linguistica Universitatis Iagellonicae Cracoviensis, 131, 299–319.

Thurman, N., & Walters, A. (2013). Live blogging: Digital journalism's pivotal platform. Digital Journalism, 1(1), 82–101.

Wahl-Jorgensen & Schmidt, T. (2019). News and storytelling. In K. Wahl-Jorgensen & T. Hanitzsch (Eds.). The handbook of journalism studies. Routledge, 261–276.

1045 - Talks 4, Syndicate 2

Chair: Diana Galvan-Sosa

- 10:45 **Michael Savage** : Corpus-driven investigation of the linguistic features of refusals in third level English language textbooks in Japan

Abstract

Refusals in Japanese are complex speech acts that serve as a decline in invitations, requests, or offers. This talk gives an overview of the social and cultural contexts which influence the use of refusals in Japanese language and how pragmatic transfer might lead to challenges when considering the bridge between 'authentic' usage and English language education. In Confucian-based settings, care is taken around the area of refusals as the surrounding linguistic and cultural factors mean that perceived transgressions can have social repercussions, endangering relationships and disrupting hierarchies. In the classroom, questions then arise about the awareness of cross-cultural pragmatic competences which students 'should' acquire and the aboutness of the textbook materials from which they come into contact with English refusal norms. This study uses the 2.3-million-word CoJET corpus of English-language textbooks from Japanese publishers used in third level settings in Japan in order to perform a systematic analysis of the structure of refusals and how English formulaic refusal patterns are structured in the textbooks. An inductive approach outlines the linguistic features of refusals as they appear in the corpus and gives an insight into how refusals are ideally used and possibly taught in the classroom. From this platform of intercultural awareness and an analysis of the CoJET corpus, findings are presented which critically consider whether the documented formulaic refusal patterns in Japanese are being adequately expressed in English language textbooks given the variety of Japanese English and its attendant cultural and linguistic factors.

- 11:05 **Shue Sum Leung, Dora Alexopoulou** : Progression and Task-based Variability of Linguistic Complexity across Grade Levels: A Case Study of the Secondary School Reading Texts in Hong Kong

Abstract

Introduction: Textbooks are vital for language learning and exam preparation. Linguistic complexity, a key component of language proficiency, is increasingly analysed using NLP tools to evaluate ESL textbooks.

Prior Studies: Previous research on ESL textbooks falls into two categories. One group compares the linguistic complexity of textbooks and exams. Another examines the progression of linguistic complexity within textbooks. However, these studies overlooked the role of progression in aligning textbooks with exams. They also failed to consider the effect of tasks (e.g. narration vs. description) on linguistic complexity.

Aims: This study examines the progression of textbooks and their comparison with exam texts, considering task effects. The research questions focus on (i) the progression of linguistic complexity across grades, (ii) the comparison of this complexity in textbooks and exams, and (iii) the influence of task types on complexity.

Data: The study inspects secondary school textbooks in Hong Kong, because they are exam-oriented early on (i.e. from grade 7). It analyses a corpus of over 400 ESL reading passages in grade 7 to 12 English textbook series and the English Language college entrance examinations taken by twelfth graders. 46 lexico-syntactic variables were extracted with Kyle's (2023) Suite of Automatic Linguistics Analysis Tools.

Findings: Rank-based ANOVA showed that (i) textbook series studied exhibit partial progression, with lexical sophistication increasing but syntactic complexity plateauing after grade 9; (ii) linguistic complexity in grades 10-12 textbooks matches or exceeds exam texts; (iii) descriptives were more linguistically complex/sophisticated than narratives, e.g. more subordination; however, each task type exhibited unique characteristics, e.g. narratives featured more low-frequency collocations attested in fictional texts.

Implications: The findings highlight the importance of exposing students to various task types and suggest that textbook evaluations should consider progression and task types to better align textbooks with exam requirements.

- 11:25 **Maicol Formentelli, Liviana Galiano, Maria Pavesi, Raffaele Zago** : Complexity matters in film and TV dialogue: An applied linguistics perspective

Abstract

The spread of digital devices and new media has considerably increased the availability of English in leisure activities, favouring (extensive) contact with the language outside educational settings and potentially leading to incidental L2 learning (Sockett 2014; Dressman/Sadler 2020). This calls for a description of English-language audiovisuals, major sources of rich and genuine language input learner-users are exposed to (Pavesi/Ghia 2020).

By adopting a corpus-based register-functional approach (Biber/Gray/Staples/Egbert 2022), this study aims to assess clausal and phrasal complexity in audiovisual dialogue with a view to comparing films and TV series with spontaneous conversation, the register they simulate onscreen (Quaglio 2009; Forchini 2012). The following questions are addressed: Which syntactic features at clausal and phrasal levels contribute to the expression of grammatical complexity in film and TV dialogue? How do they relate to the (extra)diegetic functions performed by these registers? How do films and TV series compare one to the other and to spontaneous conversation in grammatical complexity?

Moving from the Pavia Corpus of Film Dialogue (Pavesi 2014) and the Sydney Corpus of Television Dialogue (Bednarek 2018), major syntactic patterns of complexity, i.e. finite/non-finite dependent clauses and noun-phrase pre- and post-modification, are identified using POS-tag sequences, compared across the two corpora and discussed against corpus findings on English conversation (Biber et al. 2021).

Initial results show a high degree of clausal and phrasal complexity in film and TV dialogue, associated with register-specific functions of storytelling, characterisation and stance expression. Complement and finite relative clauses are markedly more frequent in TV series than films, hinting at greater linguistic elaboration of the former. Overall, fictional dialogue approximates spontaneous conversation for clausal complexity, while it exhibits higher phrasal complexity which may be partly explained in its scripted nature and in the use of genre-related specialised language. The findings corroborate the hypothesis that audiovisual dialogue represents ‘optimal input’ (Long 2020) for L2 learner-viewers accessing English in the wild.

References

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

Biber, Douglas, Gray, Bethany, Staples, Shelley, Egbert, Jesse, 2022, The Register-functional Approach to Grammatical Complexity: Theoretical Foundation, Descriptive Research Findings, Application. Routledge, London.

Biber, Douglas, Johansson, Stig, Leech, Geoffrey, Conrad, Susan, Finegan, Edward, 2021, Grammar of Spoken and Written English. John Benjamins, Amsterdam-Philadelphia.

Dressman, Mark, Sadler, Randall William, 2020, eds, The Handbook of Informal Language Learning, Wiley-Blackwell, Hoboken.

Forchini, Pierfranca, 2012, Movie Language Revisited. Evidence from Multi-Dimensional Analysis and Corpora. Peter Lang, Bern. Long, Michael, H., 2020, Optimal input for language learning: Genuine, simplified, elaborated, or modified elaborated?. Language Teaching, 53: 169-182.

Pavesi, Maria, 2014, The Pavia Corpus of Film Dialogue: A Means to Several Ends. In Maria Pavesi, Maicol Formentelli, Elisa Ghia (eds), The Languages of Dubbing. Mainstream Audiovisual Translation in Italy. Peter Lang, Bern, pp. 29-55.

Pavesi, Maria, Ghia, Elisa, 2020, Informal Contact with English: A Case-Study of Italian Postgraduate Students. Edizioni ETS, Pisa.

Quaglio, Paulo, 2009, Television Dialogue: the Sitcom Friends vs. Natural Conversation. John Benjamins, Amsterdam.

Sockett, Geoffrey, 2014, The Online Informal Learning of English. Palgrave MacMillan, London.

1045 - Talks 4, Syndicate 3

Chair: Gabrielle Gaudeau

- 10:45 **Duygu Candarli** : Multisemiotic moves in a multimodal corpus of student writing in higher education

Abstract

This study explores multisemiotic moves in a multimodal corpus of student writing, which remain largely unexplored in the literature, in UK higher education. Corpus studies have extensively examined the lexico-grammatical characteristics of university student writing in the literature (e.g. Nesi & Gardner, 2012; Staples et al., 2023); however, most of these studies have solely analysed text, and little is known about the characteristics of multimodal meaning-making practices in student writing. Given that writing at university has become largely multimodal in the digital age (Khabbazbashi et al., 2023), it is necessary to systematically analyse communicative functions of multimodal meaning-making patterns, which are defined as ‘multisemiotic moves’ in this study, in student writing. The present study addresses the following question: What are the characteristics of multisemiotic moves in a corpus of multimodal discipline-specific writing assignments in terms of frequency and function? The corpus consists of 150 multimodal discipline-specific writing assignments that received a passing grade and were written as part of degree programmes by students studying for master’s degrees in the UK in a range of disciplines from different first language backgrounds. The corpus of multimodal assignments was first tagged for each multimodal resource and its relationship with text based on an analytical framework informed by the systemic functional approach to multimodal discourse analysis (Fernández-Fontecha et al., 2019). Then, using a free corpus tool, AntConc 4.2.4, corpus-based analyses were conducted to determine the frequency and type of multimodal resources and multisemiotic moves. The findings reveal that multimodal resources are an integral part of meaning-making in university student writing, underscoring the importance of the analysis of multimodality in student writing. The most frequent multisemiotic moves were elaboration and extension, such as presenting data and comparisons, and the moves showed variation across the disciplines. This study provides methodological implications for future corpus studies on multimodal student writing. The implications will also be discussed for teaching and assessing multimodal writing at English-medium universities.

References

Anthony, L. (2023). AntConc (Version 4.2.4) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Fernández-Fontecha, A., O'Halloran, K. L., Tan, S., & Wignell, P. (2019). A multimodal approach to visual thinking: the scientific sketchnote. Visual Communication, 18(1), 5–29.

Khabbazbashi, N., Chan, S., & Clark, T. (2023). Towards the new construct of academic English in the digital age. ELT Journal, 77(2), 207–216.

Nesi, H., & Gardner, S. (2012). Genres across the disciplines: Student writing in higher education. Cambridge University Press.

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.

- 11:05 **Reka R. Jablonkai, Gail Forey** : Corpus-based analysis of high-stake exam papers for materials design for disciplinary literacy development

Abstract

This paper delves into the application of corpus-based approaches in investigating language use within the context of high-stakes school exams. While such approaches have been extensively employed in exploring disciplinary discourses and registers in higher education, there seems to be a noticeable gap in research concerning language use and disciplinary variation at the school level. This study aims to address this gap by reporting on an investigation into corpora derived from high-stakes school exam papers, particularly those associated with secondary education, such as GCSE, A Level. Acknowledging the call for a more comprehensive examination of literacy skills, especially in the context of high-stakes exams, the study focuses on three distinct school subjects: Geography, Biology, and History. The methodology involves the construction of three small corpora comprising past exam papers from these subjects. To pinpoint pertinent language aspects essential for disciplinary literacy development, interviews with subject teachers were conducted.

Subsequently, guided by these interview results, the corpus analysis focused on single-word and multi-word units, discipline-specific prefixes and suffixes. Building upon the outcomes of the corpus analysis, subject-specific teaching materials tailored for exam preparation are collaboratively developed with input from subject school teachers. Interview results highlight the importance of command words in exam papers as crucial for exam takers. Therefore, corpus analysis results particularly focusing on command words in the three subjects are discussed with pedagogical implications. In addition, an approach to the design of corpus-based teaching materials is presented. Conclusions are drawn about the significance of employing corpus-based pedagogies for enhancing disciplinary literacy development at the school level.

1145 - Break, Foyer

1200 - Talks 5, Plenary room

Chair: Elaine Vaughan

- 12:00 **Michael Handford** : Automating analysis of creativity: how can AI complement corpus analysis?

Abstract

The potential for generative AI to complement or even supplant corpus methods is a fascinating and perhaps unnerving topic (Crosthwaite and Baisa, 2023; Zappavigna, 2023). This study considers the way AI can be used to analyse creativity, an analytically challenging phenomenon, in a corpus of workplace interactions. Handford and Koester (2024) in our corpus-assisted book-length study of creative workplace discourse, distinguish between two ‘schools’ of creativity in linguistics: linguistic creativity (Carter, 2016) and discursive creativity (Jones, 2010). Whereas linguistic creativity covers the use of features like metaphors and hyperbole, discursive creativity concerns the use of language (which may not be creative in itself) to do creative things, for instance create a new identity or make an innovative change to practices. According to Handford and Koester, corpus tools can, to some degree, help pinpoint instances of linguistic creativity, but they are less adept at finding instances of discursive creativity, which is unsurprising given the highly contextual nature of this type of creativity. In terms of analysis, we propose that ‘until algorithms become effective at unearthing potentially creative uses, a qualitative analysis of the corpus is often more effective’ (2024: 594). This paper explores the extent to which this point in time has already been reached, through comparing ChatGPT and ClaudeAI’s ability to pinpoint both types of creativity in the CANBEC business-meeting corpus. I demonstrate that ClaudeAI is more effective at both analysing whole meetings, and at pinpointing instances of discursive creativity; it is, however, prone to issues of recall and precision such as hallucinations, and may conflate the analysis of different texts (which requires prior

knowledge of texts to notice). In summary, to anthropomorphise ClaudeAI, we might say it is currently a talented but slack research assistant.

- 12:20 **Mícheál J. Ó Meachair, Andrea Palandri, Gearóid Ó Cleircín** : A Synthetic Brown Corpus: Compilation, analysis, and pitfalls

Abstract

The aim of this paper is to document and engage with the significant challenges generative AI technologies present to the field of corpus linguistics, both presently and in the future. These challenges range from the core theoretical principles of the field to practical everyday issues.

We believe this topic to be pertinent in the case of all corpora, but particularly so for massive corpora where data are scraped from the web. The scale of the available web data means they cannot be manually vetted on a document-by-document basis. If vetting strategies have been applied for documents, or for entire websites, it must be noted that tests have shown human evaluators struggle to successfully distinguish synthetic texts from human-written texts (Basmov, Goldber, and Tsarfaty, 2023; Nakano, R. et al., 2021; Open AI, 2023; Tamkin, et al. 2021). Synthetic texts, in this case, refer to the output of generative AI agents. e.g. the texts that are generated as a response to a submitted prompt. Whether synthetic examples have been purposely included in a corpus or accidentally included, the fact remains that they are not evidence of bone fide language use. We believe this to be a growing issue for both wider-spoken and minority languages, but our case study focuses on the English-language because it is in English that these technologies currently perform best.

In order to explore and discuss these challenges in the context of corpus linguistics we have compiled a one-million word corpus using prompts that have been based on design specifications for the Brown Corpus (Kucera and Francis, 1967. Lenhart and Tong, 2003). We call this corpus "A Synthetic Brown Corpus" (ASBC_000001). We purposely used the indefinite article and a sequence of numbers in the codename of this corpus because the same prompt will not necessarily produce the same output, and because we believe this experiment could be re-produced at scale. Multiple widely-available generative AI technologies were used to generate the text samples to mitigate for the perceived strengths and weaknesses of one AI technology over another. Once compiled, comparative analyses of word and n-gram lists were conducted on ASBC_000001 and the original Brown Corpus as well as a collocation analysis of select features. We have purposely chosen an older English-language corpus because it made it easier for us to identify hallucinations and inaccuracies in the synthetic data, e.g. where the AI agent was instructed to write a political news article published in the USA in the year 1961, and it only named a small set of famous persons from the period. The synthetic samples we examined frequently lacked diversity and nuance in this regard, as was borne out in our empirical tests.

Basmov, V., Goldberg, Y., Tsarfaty, R. (2023) ChatGPT and Simple Linguistic Inferences:

Blind Spots and Blinds. URL: https://arxiv.org/abs/2305.14785

Kucera, H. and Francis, W. N. (1967) Computational Analysis of Present-Day American English. Brown University Press, Providence, RI.

Lenhart, S. and Tong, M. (2003) “Extracting and evaluating general world knowledge from the Brown Corpus”, in Proceedings of the HLT-NAACL 2003 Workshop on Text Meaning. URL: https://aclanthology.org/W03-0902, pages, 7-13

Open AI (2023) GPT-4 Technical Report. URL: https://cdn.openai.com/papers/gpt-4.pdf

Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., Jiang, X., Cobbe, K., Eloundou, T., Krueger, G., Button, K., Knight, M., Chess, B., Schulman, J. (2021) WebGPT: Browser-assisted question-answering with human feedback. URL: https://arxiv.org/abs/2112.09332

Tamkin, A., Brundage, M., Clark, J., Ganguli, D. (2021) Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models. URL: https://arxiv.org/abs/2102.02503

- 12:40 **Almut Köster** : Pragmatic markers and professional practice in care home interactions

Abstract

This paper reports findings from a study of care home interactions and shows the results from an analysis of frequently occurring chunks (n-grams) functioning as pragmatic markers (such as do you want, a little bit) in the interactions. The corpus of approximately 50,000 words was compiled from over 70 hours of audio-recorded naturally-occurring interactions in a care home in England. Interactions in care homes are strongly constrained by the nature of the work, where care workers interact with residents and other co-workers as they do their rounds. These interactions are typically brief and are characterized by short turns and by language that frequently accompanies actions, such as administering medication, or feeding, dressing and washing residents. Studying frequently occurring chunks can provide valuable insights into the ways in which care workers discursively construct their work and navigate the dual goals of performing their care duties and engaging interpersonally with residents. The study looks at examines the types of markers found (discourse markers vs. interactional markers as well as their specific functions, for example hedging or requesting (O’Keeffe et al., 2007; Handford, 2010). One key finding is that different chunks were used in two core areas of the professional practice of the carers: in interacting with residents on the one hand, and in talking to co-workers on the other. It is argued that these pragmatic markers index specific discursive practices within the different micro-contexts of interaction and ultimately within the macro-context of the community of practice.

References

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

O'Keeffe, A. McCarthy, M. and Carter, R. (2007). From Corpus to Classroom: Language Use and Language Teaching. Cambridge, Cambridge University Press.

1200 - Talks 5, Syndicate 1

Chair: Dawn Knight

- 12:00 **Marcel Kückelhaus** : Deciphering German AI Narratives Using Corpus Linguistics

Abstract

Artificial Intelligence (AI) is playing an increasingly prominent role in society. While linguistics has played an integral role in the development of new language-based systems, I argue that corpus linguistics has an underappreciated role to play and that is: studying AI Narratives. Several researchers have been investigating how we talk about AI in media, science, and fiction, lending insight into public opinion towards its risks and benefits, which can in turn influence the development, deployment, and regulation of AI technologies (Cave/Dihal 2023). Corpus linguistics can contribute in two important ways to this effort: 1. By providing a wealth of qualitative and quantitative methods for studying linguistic discourse, and 2. By facilitating a cross-cultural and cross-linguistic approach to the study of (global) AI Narratives.

In this presentation, I present a hermeneutic approach using quantitative and qualitative methods from corpus linguistics that can assist us in understanding public perception of AI, including how several myths and misconceptions are amplified by traditional media such as newspapers. Using CQPweb (Hardie 2012), a web-based corpus analysis tool, I approach the question of how AI is depicted in public German discourse and if that discourse is creating a modern AI mythology. This research is based on a corpus including four major German newspapers containing 4,000 articles that include the word "künstliche Intelligenz" (artificial intelligence). Those results will be compared with a smaller corpus of 1,500 tweets by German politicians or political organisations referring to "ChatGPT" or "OpenAI". By identifying narratives represented through linguistic patterns we can assess public opinion towards various contemporary AI technologies, compare them to the anglophone discourse, and determine which narratives transcend national boundaries. This work demonstrates how corpus linguistics can play an integral role in facilitating cross-cultural, cross-linguistic studies of AI Narratives.

Literature:

Cave, Stephen/Dihal, Kanta (Ed.) (2023): Imagining AI. How the World Sees Intelligent Machines. Oxford/New York: Oxford University Press.

Hardie, A. (2012): CQPweb - combining power, flexibility and usability in a corpus analysis tool. In: International Journal of Corpus Linguistics 17 (3), 380–409.

- 12:20 **Muhammad Afzaal, Xiao Shanshan** : ChatGPT OpenAI and Human Interaction: A Contrastive Parallel Corpus-based Investigation of Syntactic Complexity features in human and Machine Translations

Abstract

The rise of the artificial intelligence (AI) has generated a need for fast online translations, which human translators are unable to meet. An automatic translation from one written language to another is provided by statistically based tools such as Chat GPT, Google and Baidu Translate. This study reports the descriptive comparison of the machine-translation (MT) with human translation (HT) taking into consideration the syntactic complexity features and lexical diversity. The study uses a parallel corpus consisting of 79 texts translated from Chinese to English by professional human translators and machine translates (Chat GPT, Baidu translate & Google translate) and a comparable reference corpus of non-translated English text. The study employs Syntactic Complexity Features (Lu, 2010) for the comparison of the corpora which includes the account of fourteen features of syntactic complexity, including five dimensions, length of production units, amount of coordination, amounts of subordination, degree of phrasal sophistication and overall sentence complexity. The study shows the outcome of the machine translation and human translation in terms of the complexity, lexical diversity, and simplification in the translated texts. We show that this may be explained by the difference in lexical diversity between machine translation and human translation. Automatic metrics that measure the departure of machine translation from human translation might confuse difference with quality. This is because human and machine translations are measured using different metrics. As a result of this, the study highlights that the distinction in lexical variation that exists between machine translation and human translation receives a greater amount of attention when evaluating machine translation.

- 12:40 **Andreas Weilinghoff** : Evaluating Whisper for Sociolinguistic Data Transcription

Abstract

The transcription of sound data is an essential yet time-consuming and labour-intensive part of almost all linguistic research projects. This is especially true for corpus linguistics, as only well-transcribed and carefully annotated corpora provide a reliable basis for subsequent analyses. As recent years have seen great advancements in the field of Automatic Speech Recognition (ASR) (Jurafsky and Martin 2023), a central question is how the latest ASR models can be effectively used to enhance corpus transcription. More specifically, how do the latest ASR models perform in terms of accuracy and speed when compared to human transcribers?

This study will address this fundamental question. I will focus on how the end-to-end ASR system OpenAI Whisper (Radford et al. 2022) performs on sociolinguistic datasets. For this purpose, the spoken components of the corpora ICE Nigeria (Wunder et al. 2008) and ICE Scotland (Schützler et al. 2017) are re-transcribed with different Whisper models and the resulting transcriptions are then compared to the manual reference transcriptions via Word Error Rate (WER) metrics. To find out what significantly influences the performance of Whisper, the analysis applies linear mixed effects modelling of WER with the lme4 (Bates et al. 2015) and lmerTest (Kuznetsova et al. 2017) packages in R (R Core Team 2023). The individual files and speakers are treated as random factors.

The findings show that Whisper performs well on both varieties with significant interactions between variety and recording quality, which indicates a worse performance on ICE Nigeria due to the overall poorer quality of recordings. A crucial issue is that Whisper automatically deletes hesitations, repetitions and interruptions which can pose a challenge for sociolinguistic data transcription. Based on the findings, I will discuss key opportunities and challenges of implementing Whisper for sociolinguistic data preparation workflows.

References:

Bates, D., Maelcher, M., Bolker, B. & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1–48.

Jurafsky, D. & Martin, J. H. (2023). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. https://web.stanford.edu/~jurafsky/slp3/

Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. Journal of Statistical Software, 82(13), 1–26.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C. & Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision. https://arxiv.org/abs/2212.04356

R Core Team. (2023). R: A language and environment for statistical computing. https://www.R-project.org/

Schützler O., Gut U. & Fuchs R. (2017). New perspectives on Scottish Standard English. Introducing the Scottish component of the International Corpus of English. In Hancil, S., Beal, J. (Eds.), Perspectives on Northern Englishes (pp. 273–301). Berlin: De Gruyter Mouton.

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

1200 - Talks 5, Syndicate 2

Chair: Geraldine Mark

- 12:00 **Ana Eugenia Sancho Ortiz** : Methodological challenges in working with digitally mediated data: The compilation of the SciDis database

Abstract

The development of technology has prompted the exploration of innovative corpus-compilation techniques that enable the creation of multimodal mega- and micro-corpora of off-line and online texts (O'Keeffe and McCarthy 2022). This possibility to engage with extensive datasets entails the recognition of corpus building as an ongoing decision-making process characterized by the constant emergence of methodological challenges (Collins 2019). In this context, the SciDis (Science Dissemination) data has emerged as a static collection of digitally mediated texts aimed to represent diverse discursive phenomena within the field of scientific communication online. This database constitutes the object of study of the SciDis project, interested in the study of digital professional practices in English in the context of science dissemination, primarily characterized by a reliance on knowledge recontextualization.

This study addresses the methodological challenges in the compilation of the SciDis database, with a central emphasis on representativeness as regards digital discursive practices. The initial challenges relate to the typology of the practices considered and the need for the database to encompass the dynamic nature of digitally mediated texts. The decision was taken to observe web-hosted practices, on the one hand, and social media practices, on the other. Other difficulties were encountered regarding who generates the content and the degree to which expert users intervene in the communication of knowledge. Here, it was determined to classify the selected texts into two categories: author-generated knowledge, for the practices endorsed by users, and writer-mediated knowledge, for those wherein users mediate between authors and the knowledge they generate. Lastly, the final set of challenges pertains to the disciplinary and idiosyncratic differences identified between the three fields of knowledge selected for the analysis (health, economy and natural sciences). Thus, to ensure representativeness (Biber et al. 1998), it was decided to explore all practices specific to and shared between the disciplines but only compile and analyze those common to all. The ultimate aim of the compilation and analysis of this database is the exploration of discursive

processes that take place in digital knowledge dissemination such as recontextualization, dialogicity and identity construction.

References:

Biber, Douglas, Susan Conrad and Randi Reppen. 1998. Corpus Linguistics: Investigating Language Structure and Use. Cambridge University Press.

Collins, Luke Curtis. 2019. Corpus Linguistics for Online Communication: A Guide for Research. Routledge.

O'Keeffe, A., & McCarthy, M.J. (Eds.). (2022). The Routledge Handbook of Corpus Linguistics (2nd ed.). Routledge. https://doi.org/10.4324/9780367076399

- 12:20 **Johnatan E. Bonilla, Laura M. Merino Hernández, Miriam Bouzouita** : The pluralization of the existential verb haber 'there is/are' in written and recorded parliamentary speeches in Canarian Spanish

Abstract

The pluralization of the existential verb haber 'there is/are' has been extensively analyzed across the Spanish-speaking world (e.g., Claes 2016; Lastra & Martín Butragueño 2016; Díaz-Campos 2003). With a plural noun, the prescriptive norm dictates that the singular form must be used (había niños 'lit. there was kids'), whereby niños is considered the object of the verb. However, its pluralization is also found, especially in oral speech (habían niños 'there were kids'), where the noun appears to act as the subject, thus prompting plural agreement with the verb. This phenomenon has been identified as a feature of Canarian Spanish (e.g., ALEICan: Alvar 1975-1978; Hernández Cabrera 2016; Bouzouita & Pato 2019). As such, the objective of this presentation is to analyze the linguistic behavior of this verb in parliamentary speeches from the Canary Islands to determine whether its use has permeated even in the most formal registers.

We compare data from parliamentary speeches from (a) written parliamentary speeches from the Canary Islands' Parliament website, and (b) more than 5,300 YouTube oral recordings of parliamentary speeches found on institutional channels. YouTube's API was used to extract the automatically generated captions from these videos, leveraging YouTube's speech recognition technology to obtain caption text and timing information. Our results show that the pluralization of the verb haber is documented in both written and oral speech. However, its frequency decreases in written discourse due to the academic standard writing pressures that regulate the language, whereas the immediacy of oral speech prevents speakers from strictly following academic standards. By comparing written and oral data from the same register our results show that the corpus used influences the results and conclusions of any given study, especially if we are dealing with phenomena that goes against the linguistic norm.

Bibliography

ALEICan = Alvar, Manuel (1975-1978) Atlas lingüístico y etnográfico de las Islas Canarias, 3 vols., Madrid: La Muralla.

Bouzouita, M., & Pato, E. (2019). Antes había (n) pozos en el pueblo: la pluralización del verbo" haber" existencial en el español rural europeo. Revue de linguistique romane, 83(329), 137-165.

Claes, J. (2016). Cognitive, Social, and Individual Constraints on Linguistic Variation: A Case Study of Presentational 'Haber' Pluralization in Caribbean Spanish (Vol. 60). Walter de Gruyter GmbH & Co KG.

Díaz-Campos, M. (2003). The pluralization of haber in Venezuelan Spanish: A sociolinguistic change in real time. IULC Working Papers, 3(1).

Hernández Cabrera, Clara (2016) «Variación de haber impersonal en el español de Las Palmas de Gran Canaria». Estudios de lingüística. 30. Universidad de Alicante (ELUA)., 141-162.

Lastra, Y., & Martín Butragueño, P. (2016). La concordancia de haber existencial en la Ciudad de México. Boletín de filología, 51(2), 121-145.

- 12:40 **Yating Tao** : Decoding valency patterns and semantic senses: A comparative corpus analysis of TAKE in New Englishes and Learner Englishes

Abstract

The basic assumption of valency theory is that the verb takes a central position in a sentence as it determines how other elements combine with it to form a grammatical sentence (Herbst et al., 2004, p. xxiv). While valency theory has been applied to native English, little attention has been paid so far to non-native English varieties. The present study addresses this gap by investigating the valency patterns of two types of non-native Englishes: New Englishes (NEs) (e.g. English used in Singapore) and Learner Englishes (LEs) (e.g. English used in China), which have garnered great attention within the realms of Contact Linguistics and Second Language Acquisition research, respectively. Despite the obvious links between NEs and LEs, including shared non-native status, certain linguistic features (e.g. Nesselhauf, 2009), and common psycholinguistic processes of second language acquisition (e.g. Percillier, 2016), studies comparing the two types of varieties have remained surprisingly scarce until recently.

In this context, the present study seeks to deepen our understanding of the relationship between NEs and LEs by investigating the valency patterns of TAKE across three corpora: student writing samples from the Hong Kong and Singapore component of the International Corpus of English (ICE) and Chinese student essays from the International Corpus of

Learner English (ICLE), with the Louvain Corpus of Native English Essays (LOCNESS) as a reference. Relying on valency theory, this study aims to uncover the syntactic and semantic profiles of TAKE as well as their interplay in the two types of English varieties. Syntactically, all instances of TAKE are annotated according to the Valency Dictionary of English (Herbst et al., 2004). Semantically, the senses of TAKE distributed across different valency patterns are annotated drawing on contextual information and referring to Gilquin's (2008) semantic classification of TAKE. Finally, the relationship between the valency patterns and the senses of TAKE is discussed.

References

Gilquin, G. (2008). What You Think Ain't What You Get: Highly polysemous verbs in mind and language. In: Lapaire. J.R., Desagulier G., Guignard J.B. (Eds.), Du fait grammatical au fait cognitif. From Gram to Mind: Grammar as Cognition (pp. 235-255). Bordeaux: Presses Universitaires de Bordeaux.

Gilquin, G., & Granger, S. (2011). From EFL to ESL: Evidence from the International Corpus of Learner English. In Mukherjee J., Hundt M. (Eds.), Exploring second language varieties of English and learner Englishes: Bridging a paradigm gap (pp. 55-78). Amsterdam: John Benjamins.

Herbst, T., Heath, D., Roe, I. & Götz, D. (2004). A Valency Dictionary of English: A Corpus-Based Analysis of the Complementation Patterns of English Verbs, Nouns and Adjectives. New York/Berlin: Mouton de Gruyter.

Percillier, M. (2016). World Englishes and Second Language Acquisition: Insights from Southeast Asian Englishes. Amsterdam: John Benjamins.

Nesselhauf, N. (2009). Co-selection phenomena across New Englishes: Parallels (and differences) to foreign learner varieties. English World-Wide, 30(1), 1-25.

1200 - Talks 5, Syndicate 3

Chair: Andrew Caines (pre-recorded talks)

- 12:00 **Paul Thomas Johnson** : A New Kind of Graded Reader: The Digital Roleplaying Game

Abstract

This presentation introduces a new kind of graded reader (GR): The digital roleplaying game. Games are linguistically taxing. The grammar, vocabulary, and syntactical structures of authentic input often far exceed the abilities of language learners (Krashen, 2009). Instructors have long had this same problem with other types of authentic texts. One time-tested solution is the process of leveling, or the systematic simplification of language, to write GRs, books which are designed for students at a particular level (Waring & Takahashi, 2000). They are often a part of a series and follow a schema designed to take the learner from their current level to a target level (Nation & Wang, 1999). Extensive Reading (ER), in which students read large amounts for pleasure, with a focus on meaning, rather than directly on form, is the pedagogical foundation of the GR approach (ERF, n.d.; Krashen, 2011; Waring & Takahashi, 2000). However, it is extremely difficult to compare GRs from different series, and there is no currently accepted procedure for doing so. This is caused by a lack of transparency by publishers in the writing process, word lists used, and the corpora on which those lists are based (Robb, 2020). It is argued that a new schema based on the New General Service high-frequency word list (NGSL) solves this problem. The new, NGSL-based schema can be used to conduct cross comparison studies of previously published GRs, level institutional GR libraries, and create new, leveled materials, including the creation of a new kind of GR: The digital roleplaying game (Johnson, 2023). Finally, "Just the Facts," a leveled single-player, digital role-playing game for English Language learners levels A1-B1, is introduced. It is written using high-frequency words. For example, the first 500 words from the NGSL cover +90% of the in-game dialogue for the A1 region (iterative process education, 2024).

References:

(1) ERF. (n.d.). Graded Readers. The Extensive Reading Foundation. https://erfoundation.org/w readers/

(2) iterative process education. (2024). Just the Facts. Just the Facts. https://sites.google.com/view/ie-edu/just-the-facts

(3) Johnson, P. (2023). A New Kind of Graded Reader: The Digital Roleplaying Game. [Unpublished Master's Dissertation]. University of St. Andrews.

(4) Krashen, S. (2009). Principles and Practice in Second Language Acquisition (internet ed.). Pergamon.

(5) Krashen, S. (2011). Free Voluntary Reading. Libraries Unlimited.

(6) Nation, P., & Wang, K. (1999). Graded Readers and Vocabulary. Reading in a Foreign Language, 12(2), 355–380.

(7) Robb, T. (2020). Contradictory info on Graded Reader levels. Which do we believe? https://www.youtube.com/watch?v=giUOrbKBpSU

(8) Waring, R., & Takahashi, S. (2000). The Oxford University Press Guide to the 'Why and 'How' of Using Graded Readers. Oxford University Press Japan. http://www.robwaring.org/er/articles/Guide_to_Graded_Readers_e.pdf

- 12:20 **Xiao Liu** : Does the alignment quality of bilingual corpus have a greater impact on SMT or NMT? Taking the Spanish-Chinese legal bilingual corpus as an example, and analyzing the corpus particularities of the Spanish-Chinese language pair

Abstract

In this article a Spanish-Chinese parallel corpus in the legal field, a Chinese monolingual corpus in the legal field, and a Spanish-Chinese legal termbase will be created. Using these linguistic resources, statistical machine translation (SMT) engines and neural machine translation (NMT) engines will be trained, with translation direction from Spanish to Chinese in KantanMT platform.

Text of the Spanish-Chinese legal parallel corpus:

1) Six Spanish and EU laws that have been translated into Chinese
2) Catalan government website
3) International arbitration documents in seven European arbitration centers
4) Documents of 14 international organizations dealing with legal matters

Text of the Chinese legal monolingual corpus:

1) 24 China's laws
2) Website texts of three arbitration committees in China
3) Chinese part of the bilingual legal corpus of the 14 international organizations

A great part of the bilingual documents are aligned automatically using Hunalign. According to the alignment quality level scores, pyhton programs will be written to select the bilingual segments with scores $\geq$ 0.5, 0.6, 0.7, and train SMT and NMT engines separately. What follows is to compare the automatic evaluation metrics of the machine translation engines trained by KantanMT: F-measure, BLEU and TER (To be clearer, independent variable: aligment score $\geq$ 0.5, 0.6, 0.7; dependent variable: corresponding automatic evaluation metrics F-measure, BLEU and TER). The research hypothesis is that the alignment quality of bilingual corpus has a greater impact on SMT results than NMT results.

In addition to the automatically aligned bilingual corpus, there are also a lot of manually aligned bilingual corpus in this study to meet the corpus size required to launch machine

translation. Both manually aligned and automatically aligned corpus will be managed together. When applying regular expressions to manage the corpus, it has been discovered that Chinese can be used as an interlanguage to separate the mixed English and Spanish texts. Furthermore, the characteristics of Spanish and Chinese language pairs will be analysed when applying regular expressions and doing alignment. For example, the difference between Chinese full-width and Spanish half-width punctuation marks has a great impact on the quality of automatic alignment. The next step is to explore an innovative way to divide Spanish and Chinese terms written together into two columns automatically. And the last step is to study how to batch convert files in different formats, for example, convert tab-delimited txt to tmx and xlsx.

- 12:40 **Jean Marguerite Jimenez, Ida Ruffolo** : FAIR AND FAST FASHION? A content and linguistic analysis of gender equality in CSR reports

Abstract

The 2022 World Economic Forum's Global Gender Report affirms that gender parity may not be achieved for another 132 years. Although this represents a slight improvement compared to the 2021 estimate, there is still a long road ahead. Indeed, "women's workforce outcomes are suffering and the risk of global gender parity backsliding further intensifies," cites the report.

Through the 2030 Sustainable Agenda, the United Nations has made a strong call towards meeting 17 Goals, one of which (#5) is dedicated to achieving gender equality and empowering all women and girls (United Nations, 2019). Over 100 countries have committed to this Agenda and corporations worldwide are creating strategies which strive for a more equal world. Specifically, companies are addressing gender issues in the context of Corporate Social Responsibility (CSR) and sustainable development through policies and initiatives promoting gender equality. Yet, the fast fashion industry, which targets mainly the female population, faces major challenges related to women's wellbeing and labor rights.

Effectively communicating gender-related initiatives to its stakeholders is crucial for a company to gain a financial advantage. While several studies have investigated rhetorical and discourse features of CSR reports using corpus linguistics (e.g., Catenaccio 2013; Aiezza 2015; Bondi 2016; Fuoli 2018; Crawford Camiciottoli, forthcoming), fewer studies have focused specifically on the language used to report gender issues in CSR reports (Haji, Hossain, 2016; Haynes, 2017; Hossain et al., 2017).

Based on these premises, this paper discusses how 2 leading fast fashion companies, Inditex and GAP Inc., relay gender equality issues in their CSR reports as part of their communication strategy. The study adopts a corpus-based discourse analysis approach to identify the topics related to gender equality included in these companies' CSR reports and the linguistic strategies employed to communicate their values and impacts on economic, social and environmentally sustainable development.

REFERENCES

Aiezza, M.C. 2015. BRIC by BRIC. The CorSus of Corporate Social Responsibility. A Corpus-Assisted Discourse Analysis of Sustainability Reports, Journal: Università degli Studi di Napoli Federico II.

Bondi, M. 2016. The future in reports. Prediction, commitment and legitimization in corporate social responsibility (CSR). Pragmatics and Society, 7(1), 57 - 81

Catenaccio, P. 2013. The discursive encoding of changing business values in CSR reports: a corpus-based investigation. In: F. Poppi, W. Cheng (eds.) The three waves of globalization: winds of change in professional, institutional and academic genres. Newcastle upon Tyne: Cambridge Scholars, 56-76.

Crawford Camiciottoli, B. forthcoming. Exploring the language of transparency: A comparative analysis of the sustainability reports of U.S. vs. Italian fashion brands. ESP across Cultures.

Fuoli, M. 2018. Building a Trustworthy Corporate Identity: A Corpus-Based Analysis of Stance in Annual and Corporate Social Responsibility Reports. Applied Linguistics, 39(6), 846-885.

Haji, A.A. and Hossain, D.M. 2016. Exploring the implications of integrated reporting on organizational reporting practice: evidences from highly regarded integrated reporters. Qualitative Research in Accounting and Management, 13(4), 415-444.

Haynes, K. 2017. Accounting as gendering and gendered: a review of 25 years of critical accounting research on gender. Critical Perspectives on Accounting, 43, 110-124.

Hossain, D.M., Ahmad, N.N.N. and Siraj, S.A. 2017. Power relationships in gender-related disclosures: exploring language in selected fortune 500 companies' sustainability reports. International Journal of Business Governance and Ethics, 12(3), 262-288.

United Nations, 2019. Sustainable Development Goals, available at: https://www.un.org/sustainabledevelopment/.

World Economic Forum, 2022. Global Gender Gap Report 2022, available at: https://www3.weforum.org/docs/WEF_GGGR_2022.pdf

1300 - Lunch, Foyer

1400 - Talks 6, Plenary room

Chair: Christopher Fitzgerald

- 14:00 **Diane Nicholls, Andrew Caines, Paula Buttery** : The Write & Improve Corpus 2024

Abstract

We present a new annotated corpus of written learner English, derived from essays submitted to the learning platform Write & Improve (W&I). Users of W&I are presented with automated scoring and feedback on grammatical errors, and are encouraged to act on their error feedback, submitting multiple versions of their essays for any given prompt. We build the corpus on this interplay between *users* and *prompts*, collecting sets of essays submitted by users for a restricted list of 50 popular prompts. The prompts include 20 aimed at beginner learners of English, 20 aimed at intermediate learners, and 10 at advanced learners. This distribution reflects the greater use of W&I by beginners and intermediate learners of English. We ensured that the prompts were not likely to elicit personal information, and covered a broad range of tasks and topics. This list of prompts enabled us to identify 5000 essay sets written by 800 users, forming the basis for the Write & Improve Corpus, which will be released this year for non-commercial use by Cambridge University Press & Assessment. We describe the steps we took to ensure the corpus contains appropriate texts, does not include personal information, and comes with annotations relating to CEFR level and grammatical errors. All essays were submitted between 2020 and 2022 by registered users of W&I who have supplied their first language (L1) in an optional questionnaire. In total, there are more than 23,000 essays containing more than 3 million word tokens. There are 23 different L1s in the corpus, with the most common being Spanish, Portuguese, Japanese, Arabic and Vietnamese. The Write & Improve Corpus 2024 is scheduled for release this Autumn: we will give details about how to stay informed about the release of the corpus.

- 14:20 **Saara Hellström** : Comparing French and Swedish web registers using multilingual word vectors

Abstract

The web features a wide variety of registers (Biber, 1988), i.e., situationally defined language use with different purposes (e.g., blogs, recipes), in numerous languages. Yet online language use in other languages than English (Biber & Egbert, 2018) remains largely unexplored, since we lack register-annotated web corpora. Moreover, cross-linguistic analyses are usually manually conducted which is time-consuming and prone for subjective interpretations. Our study expands web register research to French and Swedish and examines the register characteristics using multilingual word vectors allowing the analysis of registers in one multilingual space without manual comparison.

Our data consists of the newly established FreCORE and SweCORE including similarly register-annotated web documents. In our analysis, we first extract the keywords, i.e., the statistically overrepresented words indicating 'aboutness' of the texts, from the corpora using text dispersion keyness to get the language specific characteristics for the registers. Then, using the fastText tools, we transform the keywords into word vectors, i.e., linguistically motivated, numerical representations of words derived from a language model. The word vectors present words in one multilingual space where semantically similar words are represented by similar vectors even across languages. Finally, to examine the cross-lingual similarities of the keywords and what they tell about the registers, we cluster the word vectors with KMeans.

Twenty clusters offer the best fit to the data. Our analysis shows that the clusters group keywords based on their topical or grammatical features: e.g., cluster 1 (topic) includes *politique* - *politik* (politics), *peuple* - *folket* (people) while cluster 10 (grammar: stance) features *pense* - *taenker* (think), *vrai* - *sant* (true). Nearly all clusters include French and Swedish keywords sharing the semantic meaning which indicates cross-linguistic similarities in registers. The keywords within a register group coherently which suggests that clustering could be a viable method to group keywords computationally instead of the laborious manual grouping.

Biber, Douglas. 1988. Variation Across Speech and Writing. Cambridge: Cambridge University Press.

Biber, Douglas & Jesse Egbert. 2018. Register Variation Online. Cambridge: Cambridge University Press.

- 14:40 **Ivor Timmis** : ‘Dear Sir, write to mee on Mr Mudge’s backside’. Resourcefulness in letters by the poor and desperate, c.1760-1830

Abstract

This talk draws on three small historical corpora of letters written by the poor and desperate c.1760-1830: the Pauper Letter Corpus, comprising letters written to the parish overseers asking for charitable relief; the Prison Letter Corpus, comprising letters written by women facing deportation to Australia asking for a degree of clemency, and the Threatening Letter Corpus, comprising colourful anonymous letters to the wealthy demanding relief from poverty on pain of death. One of the most impressive features of all three sets of letter is the way the writers, usually with low levels of education and literacy, were able to muster their linguistic resources for a writing task crucial to their welfare. They showed great resourcefulness, I will argue, in three main ways: by repurposing formulaic language from texts, such as the bible and popular stories, with which they were familiar; by exploiting cultural motifs of the time such as the ‘deserving poor’and, particularly in the case of the threatening letters, by displaying a certain mischievous creativity. These letters, I will conclude, shed light on the nature of literacy at the time, suggesting that in this context it can be seen as a communal asset as much as an individual accomplishment. We will also discuss whether the three corpora illustrate what Carter (2004) has called ‘demotic creativity’.

Reference

Carter, R. (2004) Language and Creativity: the art of common talk. London: Routledge.

1400 - Talks 6, Syndicate 1

Chair: Gabrielle Gaudeau

- 14:00 **Yahui Wang** : UK Media representation of COVID-during the first national lockdown: a corpus-assisted critical discourse analysis

Abstract

News media is understood as the channels that carry messages to numerous audiences, and news media cannot report on all the events taking place across the world, there must be selection criteria that operate in order for an event to be considered news. Therefore, the media has the power to frame a social event based on their preference and influence over its audiences, shaping public ideas and behaviours (Bednarek & Caple, 2012). Data was collected from the Sun and Guardian, and analysed quantitatively and qualitatively by applying corpus linguistics and critical discourse analysis combining the representation of social actors (Van Leeuwen, 2008) and discursive strategy (Reisigl & Wodak, 2001). Keywords and collocates of corpus linguistics reveal the frequently used metaphors and evaluation to construct COVID-19, related mitigations and social groups. Governmental officials are highlighted by collocating with reporting verbs. Further concordance analysis within the predication strategy shows that the coronavirus and mitigation are negatively evaluated, and healthcare workers are constructed as heroes who fight against the virus on the frontline. Additionally, the construction of workers from other fields and family members are also examined. According to the representation of social actors, the news stories improve the credibility of governmental officials through collectivization and aggregation strategies. Finally, the linguistic findings are discussed through the sociopolitical context as well as journalistic criteria based on the news values (Atanasova & Koteyko).

Reference

Atanasova, D., & Koteyko, N. (2017). Obesity frames and counter-frames in British and German online newspapers. Health, 21(6), 650-669.

Bednarek, M., & Caple, H. (2012). News discourse. A&C Black.

Reisigl, M. & Wodak R. (2001) Discourse and discrimination: Rhetorics racism and anti-semitism. London: Routledge.

Van Leeuwen, T. (2008) Discourse and Practice. Oxford: Oxford University Press.

- 14:20 **Justin McNamara, Michaela Rusch** : Only a Matter of Words-A Corpus-Based Study of the Websites of Popular European Businesses Analysing Diversity and Inclusion Notes

Abstract

Established companies are fully aware of their position in the market, their role in society, and what attributes to succeed as a business. However, in the current market, success has become, more often than not, a challenge. As diversity becomes one of the major aspects of company culture and communication, the question arises as to how it is employed further. Of late, there has been a sharp increase in images of different ethnicity, racial and gender identities, sexual orientation and socioeconomic status groups in the advertising of a multitude of companies in Europe and worldwide. This corpus-based study will investigate the genre of diversity sections on the websites of European companies using a random selection of examples from Eire, the UK, Germany, France and Spain. Investigating those texts, this study will focus on the visual, linguistic and multimodal elements used by the companies to identify them as diverse, inclusive and perhaps linguistically sensitive. The researchers will analyse company statements, social media posts, diversity reports and overall determine how diverse and inclusive these companies are in their perceived company culture. As previous studies have pointed out, communicating diversity is used as a tool, but it also can present a challenge once the company culture refrains from executing it fully. For companies, diversity notes need to manifest a vivid company culture to avoid double

standards. In this paper the authors illustrate through distinct examples that diversity and inclusion are integral tools in business communication. Customers as well as employees want to be represented and included, and doing so leads to the credibility and sustainability of the company. Up to this point, the study could – by employing the conventionally established techniques of corpus analysis – show that a distinct set of register is employed, such as in sample phrases like “diversity and inclusion”, “foster inclusion”, “foster awareness”, “foster diversity” and as in “Promoting cultural and social diversity” (SNCF).

- 14:40 **Richard Badger** : Following the science: a study of the Downing street Covid-19 press conferences

Abstract

During the Covid-19 pandemic, press conferences on behalf of the UK government from 10 Downing Street were presented by politicians and medical experts. This was one of the strategies the UK government used to demonstrate that their policies were following the science.

This study combines corpus linguistic and functional linguistic techniques to examine, first, what was distinctive about the press conference language and, second, the differences between the language used by politicians and the medical experts. The study is based on a corpus of transcripts of the Downing Street Press conferences conducted between March 3, 2020, and April 5, 2021, constituting approximately 1.2 million words.

The corpus linguistic analysis used key word analysis to identify what is distinctive about the press conferences as compared with General English usage (English web 20102), other spoken English (Spoken BNC) and Covid language (Covid Open Research Data Set). The study then explores how the politicians’ and scientists’ language differs using a key word analysis and a functional grammar process analysis. The study compares the use of the two approaches to corpus analysis.

1400 - Talks 6, Syndicate 2

Chair: Diana Galvan-Sosa

- 14:00 **Graham Burton, Maria Cristina Gatti** : The word list repository: creating an open resource for researchers and teachers

Abstract

As Timmis (2015:41) notes, 'perhaps the most obvious application of a corpus in the field of lexis is its ability produce automatically more comprehensive and reliable frequency lists than can be generated manually'. It does not seem controversial to suggest that we should consider frequency of occurrence when selecting which vocabulary we want to teach and corpora can clearly give us such information quite easily. Since the pioneering lists created before the advent of electronic corpora (e.g. Thorndike and Lorge's 1944 Teacher's Wordbook of 30,000 Words), a large number of word lists have been compiled. These include lists for general language domains (e.g. Brezina and Gablasova's 2015 'New General Service List'), of multi-word items (e.g. Martinez and Schmitt's 2012 'Phrasal Expressions List') and for specific language domains (e.g. Dang's 2018 'Hard Science Spoken Word List').

A curriculum designer or teacher looking for such word lists, or a researcher hoping to share her or his own list with teachers or the research community, immediately encounters a problem, however – no single repository for word lists exists. Some are available as appendices to journal articles, others are hosted on personal or institutional websites, while others still are hosted on platforms such as Lextutor and OSFHome. This talk presents a planned open-access, online depository we are currently creating at the Centre for Academic Writing, Free University of Bozen-Bolzano, which aims to offer a single access point for word lists, independently of field or language. After a brief overview of the characteristics of currently existing word lists, we will outline the proposed functionality of the new site. We will also outline plans for the potential future expansion of the site, for example by including teaching materials based on the word lists, or direct links to concordances. We would be very much interested in collaborating with other researchers in developing the site and hope that this talk will provide a stimulus for such work.

Brezina, V., & Gablasova, D. (2015). Is There a Core General Vocabulary? Introducing the New General Service List. Applied Linguistics, 36(1), 1–22.

Dang, T. N. Y. (2018). A Hard Science Spoken Word List. ITL - International Journal of Applied Linguistics, 169(1), 44–71.

Martinez, R., & Schmitt, N. (2012). A Phrasal Expressions List. Applied Linguistics, 33(3), 299–320.

Thorndike, E. & Lorge, I. (1944). The teacher's word book of 30,000 words. New York: Columbia University.

Timmis, I. (2015). Corpus linguistics for ELT: Research and practice. Oxford: Routledge.

- 14:20 **Joan O'Sullivan, Tamami Shimada** : Introducing the Corpus of New Speakers of Irish English

Abstract

Ireland has undergone dramatic demographic change in recent decades due to net immigration; this has led to research studies on the sociolinguistics of this change in terms of sociolinguistic variation, language ideology and identity (e.g. Migge 2012; Diskin 2016; Nestor et al. 2012; Corrigan and Diskin 2020). However, there is a dearth of studies which use corpus linguistics in analysing the relationship between immigrants to Ireland and the variety of English spoken in Ireland, Irish English (IrE). In this paper, we discuss the design, development and application potential of a new corpus of the English spoken by immigrants in Ireland. The corpus provides an important empirical dataset for researching 'new speakers' of IrE and will facilitate comparative studies with existing corpora of IrE such as ICE-Ireland (Kallen and Kirk 2008) and the Limerick Corpus of Irish English (LCIE) (Farr, Murphy and O'Keeffe 2004). In naming the corpus, we extend the concept of the 'new speaker', associated with later acquisition of a particular language, to include new speakers of varieties of a particular language, in this case IrE. The analysis of the new speaker data will take a bottom-up approach and will allow comparison with native IrE speaker data in terms of salient IrE features (grammatical, lexical, phonological, pragmatic).

References:

Corrigan, K.P. & Diskin, C. 2020. 'Northmen, southmen, comrades all'?: The adoption of discourse like by migrants North and South of the Irish border. Language in Society 49(2).

Diskin, C. 2016. 'Standard language ideologies in multicultural Ireland: A case study of Polish and Chinese migrants in Dublin', in Regan, V., Diskin, C. and Martyn, J. (eds.) Language, Identity and Migration: Voices from Transnational Speakers and Communities, 287-326. Berlin: Peter Lang.

Farr, F., Murphy, B. and O'Keefe, A., 2004. The Limerick corpus of Irish English: Design, description, and application. TEANGA, the Journal of the Irish Association for Applied Linguistics, 21, pp.5-29.

Kallen, J.C. and Kirk, J., 2007. ICE-Ireland: Local variations on global standards. In Creating and Digitizing Language Corpora: Volume 1: Synchronic Databases (pp. 121-162). London: Palgrave Macmillan UK.

Migge, B., 2012. Irish English and recent immigrants to Ireland. New Perspectives on Irish English, 44, p.311.

Nestor, N., Ní Chasaide, C. and Regan, V. 2012. 'Discourse 'like' and social identity – a case study of Poles in Ireland', in Migge, B. and Ní Chiosáin, M. (eds.) New Perspectives on Irish English, 327-354. Amsterdam/Philadelphia: John Benjamins.

- 14:40 **Anastasia Shavrina, Anastasia Vyrenkova, Sergei Obiedkov, Anastasia Chivikova** : Crowdsourcing for Error Correction in L2 Writing

Abstract

L2 learner corpora are essential for second-language research and serve as a valuable resource for training machine learning models to help design spellcheckers and tools facilitating the language learning process. Building such a corpus requires not only collecting writing samples but also identifying, correcting, and classifying errors therein. When done manually, this is time-consuming. Error extraction and classification can be performed automatically by tools such as ERRANT [1] if both original and corrected sentences are provided. Here we study the possibility of using crowdsourcing to obtain corrections.

We took over 34,000 sentences from the Russian Learner Corpus (RLC) [2] and had them corrected by users of the Toloka crowdsourcing platform [3]. With almost 5,000 users involved, each sentence was corrected by at least five users resulting in over 200,000 corrected sentences. Most of the original sentences had no corrected counterparts in RLC; however, for quality control, we included 33 sentences where 63 errors of various types had been identified and annotated. These sentences were shown to several hundred Toloka users each.

We analyzed every error across three dimensions reflecting its amenability to correction via crowdsourcing: visibility (how often users attempt to correct the error), precision (how often a correction attempt results in the expected correction), and recall (the product of the visibility and precision, the proportion of users correcting the error as expected). Based on this, we clustered errors into groups, which lead us to several hypotheses. In particular, our analysis suggests that errors resulting in non-existent forms are rarely overlooked, while syntactic and lexical errors are harder to correct. Other factors include whether a token contains a single or multiple errors and whether there are other errors in the sentence. We plan to validate the hypotheses using relevant subsets of the 34,000 sentences and in additional experiments.

References

1. Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 793–805, Vancouver, Canada. Association for Computational Linguistics.
2. Ekaterina Rakhilina, Anastasia Vyrenkova, Elmira Mustakimova, Alina Ladygina, and Ivan Smirnov. 2016. Building a learner corpus for Russian. In Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition, pages 66–75, Umeå, Sweden. LiU Electronic Press.
3. https://toloka.ai (https://toloka.ai/)

1400 - Talks 6, Syndicate 3

Chair: Yuan Gao

- 14:00 **Artur Tsymbalyuk, Lilia Shevyrdyaeva** : We used vs. We demonstrate: a diachronic comparison of self-mention in high-stakes research articles in life sciences

Abstract

The use of personal pronouns as a method of expressing authorial presence and discursive identity is one of the most visible features of scientific writing changing over time. Numerous works have focused on diachrony in the language (Li, 2021; Wang & Hu, 2023; Zhou, Gao & Lu, 2023). Particularly influential was the research into personal pronouns use across a range of scientific disciplines and genres (Hyland, 2001, Harwood, 2005, Dontcheva-Navratilova, 2023).

This study aims to examine the dynamics of self-mention in high-stakes scientific publications over prolonged periods of time. In this study we focus on biological discourse, therefore, biology-related articles from Nature were selected within the period from 1950 to 2020. Eight subcorpora were compiled representing eight points in time with a 10-year interval starting from 1950 and ending in 2020 (between 57,537 and more than 1 mln tokens each) to trace the dynamic changes in self-mention (plural 1st person personal pronouns we, us and possessive adjective our) as a powerful rhetorical strategy for emphasising a writer's contribution. Using AntConc we calculated frequency and distribution of pronouns followed by qualitative analysis of most frequent collocates.

The study shows the steadily increasing frequency of self-mention which reflects diachronic changes in the perception of writer's self-awareness in the texts and growing emphasis on personality and authorial contribution which was not typical of the language used in earlier decades when the information was presented mostly in impersonal form. The shift in rhetorical function from explaining methodology (we used / applied) to presenting findings (we found / show) demonstrates variation in self-representation of scientific writers.

Preliminary results confirm the need for further study to reveal the changes in biological academic discourse along with the driving forces, which will contribute to teaching academic writing to the students of life sciences.

Dontcheva-Navratilova O. (2023). Self-mention in L2 (Czech) learner academic discourse: Realisations, functions and distribution across master's theses. Journal of English for Academic Purposes, 64 (2023) 101272. https://doi.org/10.1016/j.jeap.2023.101272.

Harwood, N. (2005). "Nowhere has anyone attempted ... In this article I aim to do just that." Journal of Pragmatics, 37(8), 1207–1231. doi:10.1016/j.pragma.2005.01.012

Hyland, K. (2001). Humble servants of the discipline? Self-mention in research articles. English for Specific Purposes, 20(3), 207–226. doi:10.1016/s0889-4906(00)00012-0.

Li, Z. (2021). Authorial presence in research article abstracts: a diachronic investigation of the use of first person pronouns. Journal of English for Academic Purposes, 51, 10.1016/j.jeap.2021.100977.

Wang, Y., Hu, G. (2023). Shell noun phrases in scientific writing: A diachronic corpus-based study on research articles in chemical engineering. English for Specific Purposes, Volume 71, 178-190. https://doi.org/10.1016/j.esp.2023.05.001.

Zhou, X., Gao, Y. & Lu, X. (2023). Lexical complexity changes in 100 years' academic writing: evidence from nature biology letters. SSRN Electronic Journal. 10.2139/ssrn.4331103.

- 14:20 **Reem F. Alfuraih** : Competence and Creativity Indicators: A Taxonomy of Translation Positive Practices in The Undergraduate Learner Translator Corpus

Abstract

Learner translator corpora emerged at the intersection between learner corpus research and corpus-based translation studies. Like most of learner corpora, the focus of learner translation corpus research is on error analysis. The shared annotation among learner translator corpora is error annotation such as MeLLANGE (Castagnoli et al. 2011). Little attention has been given to the analysis and annotation of competence and creativity indicators in learner translation corpus research. Based on the existing models of translation competence (PACTE 2005) and (Göpferich 2009), this paper focuses on the development of a competence and creativity taxonomy of translation positive practices and the tagsets used to highlight the competence and creativity indicators in the Undergraduate Learner Translator Corpus (ULTC). One of the main objectives of the ULTC is to develop a corpus-driven quality-assessment framework that measures competence, creativity, and positive practices in the translations of undergraduate learners from and into Arabic compared to reference corpora data.

The Undergraduate Learner Translator Corpus (ULTC) is an ongoing parallel corpus of Arabic and English data produced by learners of translation. The project includes a set of reference corpora of published translations designed for the purpose of comparison. One of the main potentials of the resource is the synergies between multiple corpora that present different modules i.e. written and multimodal, and multitargets of a source text translated by one learner only i.e. draft and final of the graduation project or multitargets of the same source text translated by different learners i.e. multitarget translation of the assignments and tasks in other proficiency levels.

The proposed taxonomy focuses on empirical product- and process-oriented data translation quality, the revision process, and translator self-assessment. It highlights aspects creativity reflected in instances of novelty and problem-solving activities. It includes different levels of translation competence which are bilingual competence, pragmatic competence, profession-related competence, transfer competence, and strategic competence. The proposed taxonomy is expected to have pedagogical implications (i.e. designing tests that can measure translation competence and informing textbooks).

References

Alfuraih, R. (2019). The Undergraduate Learner Translator Corpus: A New Resource for Translation Studies and Computational Linguistics. Language Resources and Evaluation. DOI: 10.1007/s10579-019-09472-6.

Castagnoli, S., Ciobanu, D., Kunz, K., Volanschi, A., & Ku¨bler, N. (2011). Designing a learner translator corpus for training purposes. In N. Ku¨bler (Ed.), Corpora, language, teaching, and resources: From theory to practice (pp. 221–248). Bern: Peter Lang.

Göpferich, Susanne. (2009). Towards a model of translation competence and its acquisition: The longitudinal study TransComp. In Behind the mind: methods, models and results in translation process research, eds. Arnt Lykke Jakobsen, and Inger Mees, 11–37. Samfundslitteratur.

PACTE. (2005). Investigating translation competence. Meta 50 (2): 609–619.

PACTE. (2009). Results of the validation of the PACTE translation competence model: Acceptability and decision making. Across Languages and Cultures 10 (2): 207–230.

- 14:40 **Michael T. L. Pace-Sigge** : Presence and Absence of Laughter and Gestures. Examples from the BNC-Spoken 2014 and Dickens' Novels

Abstract

This paper is concerned with non-verbal discourse markers in different types of conversation. Partington (2014) highlights that corpus-linguistics appears to be primarily concerned with presence. When critically reviewing different forms of discourse this might, however, lead to the researchers turning a blind eye to the context and co-text within which these exchanges take place.

Two case studies will present the presence - or lack of it - of two extra-lingual features which provide crucial co-text markers that provide relevant pointers towards the context in which they occur in two separate corpora. Multi-modal research in linguistics, cognitive studies, and psychological studies have shown that, in spoken discourse, two key markers tend to be present in most situations: one is the use of laughter; the second is that of how gestures are employed to underscore and clarify what is being verbally transmitted. In this, the chapter revisits work done by Alan Partington (2006), who looked at 'laughter in corpora'; for this case study, the corpus in question will be the most recent comprehensive corpus of casual spoken British English, namely the BNC2014 Spoken. The second case study looks at natural use of gesture in conversations as described by Gullberg (2006). For this case study, references to gesture during conversations will be observed in a corpus of the works of Charles Dickens (Mahlberg et al, 2013), which presents an example of highly descriptive fiction.

Overall, the research findings show that a well-constructed corpus can deliver a wealth of information of when and how laughter is employed by speakers even if the phonological qualities of these vocalisations are not available. This highlights the fact that, even with the absence of audio, a well-designed corpus can still provide a wealth of extra-lexical information which can form the basis of research.

References

British National Corpus 2014: User Manual and Reference Guide, Version 1.1. (BNC2014). http://corpora.lancs.ac.uk/bnc2014/doc/BNC2014manual.pdf (last accessed 21/10/2023).

Gullberg, M. (2006). Handling discourse: Gestures, reference tracking, and communication strategies in early L2. Language learning, 56(1), 155-196.

Mahlberg, M., Smith, C., & Preston, S. (2013). Phrases in literary contexts: Patterns and distributions of suspensions in Dickens's novels. International Journal of Corpus Linguistics, 18(1), 35-56.

Partington, A. (2006). The linguistics of laughter: A corpus-assisted study of laughter-talk. Routledge.

Partington, A. (2014) Mind the gaps. The role of corpus linguistics in researching absence. International Journal of Corpus Linguistics 19 (1): 118-146.

Scott, M. (2023). WordSmith Tools version 8, Stroud: Lexical Analysis Software

1500 - Break, Foyer

1530 - Talks 7, Plenary room

Chair: Pascual Pérez-Paredes

- 15:30 **Elaine Vaughan, Brian Clancy** : Eat, pay, love: A corpus perspective on the language of first dates

Abstract

Romantic dating is a fundamental social activity, yet from a linguistic perspective research on the language of dating in natural settings is scarce, for obvious reasons. There is, however, a burgeoning interest in the value of non-scripted reality TV dating programmes to provide linguistic and pragmatic insights into these primarily 'occluded genres' (after Loudermilk 2007). First Dates is a reality TV show, which features people on blind dates set up by the show's production team. Since first being broadcast in the UK, the format has since been franchised internationally, including in the Republic of Ireland. This paper uses a corpus of interaction from the Irish version of the show, First Dates Ireland (FDIrE), and explores how the interaction in FDIrE can be analysed as a site of intimate discourse, albeit nascent. The type of interaction, a first date, is a high stakes initial encounter (see Haugh & Sinkeviciute 2021), and this is amplified by the setting and the situation, making FDIrE a rich site for the analysis of how participants negotiate complexities of this initial encounter in their use of language. While the interaction itself may be problematised from the perspective of authenticity and performance, the interaction between the potential couples is unscripted and represents a valuable and interesting example of intimate discourse (Clancy 2016). We take a corpus pragmatic approach to FDIrE exploring the pragmatic markers that characterise the context, and bringing these into relief by comparing them to naturally occurring, established intimate discourse such as that represented in existing, comparable corpora (ibid.). Findings show that the frequency of use of specific pragmatic markers are associated with particular discursive routines in this context. The argument is put forward that these markers are those responsible for the establishment, rather than the maintenance of intimacy.

Clancy, B. (2016) Investigating Intimate Discourse. Exploring the Spoken Interaction of Families, Couples and Friends. Routledge.

Haugh, M. & V. Sincaviciute (2021) The pragmatics of initial interactions: Cross-cultural and intercultural perspectives. Journal of Pragmatics, 185: 35–89.

Loudermilk, B. (2007) Occluded academic genres: An analysis of the MBA Thought Essay. Journal of English for Academic Purposes, 6(3): 190–205.

- 15:50 **Nina Haket, Ryan Daniels** : Why Making Words Better Is Not That Simple: Conceptual Engineering and Distributional Semantics

Abstract

Conceptual Engineers want to change what words mean; they are optimistic that in doing so, we can improve both society and science for the better. They find a word with a defect, and aim to provide a better alternative. However, despite having largely applied goals and wanting to impact real-world discourse, conceptual engineering is frequently still done from the armchair. Conducting or drawing upon experimental studies is starting to become more popular (see e.g. Fischer, 2020; Landes & Reuter, 2023; Machery, 2021), but corpus work is still severely under-utilised. Where it is used, it usually concerns inter-word relationships within a semantic network (Kobylarz et al., 2023). Notably absent from existing research, to the best of my knowledge, is an intra-word analysis of lexical variation. Words are not as static as engineers would like. Furthermore, despite being the most utilised form of language, no research has focused exclusively on spoken conversational data.

This work uses BERT word embeddings applied to the spoken component of the BNC14 (Love et al., 2017). Against a theoretical background of linguistic contextualism about meaning (Jaszczolt, 2015; Recanati, 2010), I explore how we actually use the terms that conceptual engineers frequently aim to target using metrics like Maximum Explained Variance and SelfSimilarity (Ethayarajh, 2019). By going beyond considering polysemy and looking at constant lexical modulation in discourse, I aim to show that the creation of a single “improved meaning”, as many conceptual engineers aim for, simply cannot be successful. Even pluralism, or the idea that we can engineer different meanings of the same word for different purposes (Belleri, 2021; Sawyer, 2021), falls short of capturing the flexibility and malleability of lexical meaning in context. This work forms part of a larger project concerning the intersection of conceptual engineering, linguistic contextualism, and distributional semantics.

Belleri, D. (2021). On Pluralism and Conceptual Engineering: Introduction and Overview. Inquiry: An Interdisciplinary Journal of Philosophy.

Fischer, E. (2020). Conceptual control: On the feasibility of conceptual engineering. Inquiry: An Interdisciplinary Journal of Philosophy, 1–29.

Jaszczolt, K. M. (2015). Default Semantics. In B. Heine & H. Narrog (Eds.), The Oxford Handbook of Linguistic Analysis (pp. 743–770). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199677078.013.0009

Landes, E., & Reuter, K. (2023, January 9). Empirical Data on the Implementation of Engineered Concepts. The New Experimental Philosophy Blog. https://xphiblog.com/empirical-data-on-the-implementation-of-engineered-concepts/

Löhr, G., Kobylarz, B., & Günther, F. (2023, December 7). Conceptual engineering must be informed by distributional semantic models [Presentation]. Workshop on Empirical Conceptual Engineering: Implementations, Challenges, and its Theoretical Basis, Department of Philosophy, University of Zürich.

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken BNC2014:

Designing and building a spoken corpus of everyday conversations. International Journal of Corpus Linguistics, 22(3), 319–344. https://doi.org/10.1075/ijcl.22.3.02lov

Machery, E. (2021). A new challenge to conceptual engineering. Inquiry, 1–24. https://doi.org/10.1080/0020174X.2021.1967190

Recanati, F. (2010). Truth-Conditional Pragmatics. Oxford University Press.

Sawyer, S. (2021). Concept Pluralism in Conceptual Engineering. Inquiry: An Interdisciplinary Journal of Philosophy, 1.

- 16:10 **Isolde van Dorst, Mathew Gillings, Jonathan Culpeper** : Impoliteness variation in Britain: a corpus-based study

Abstract

Building upon work within variational pragmatics (e.g., Schneider and Barron, 2008), this paper explores the scope of impoliteness variation in Britain. Through the use of corpus-based methods, we examine the frequencies of different impoliteness structures in the Spoken British National Corpus 2014 (Love et al., 2017), an approximately 11.5 million word corpus of everyday spoken British English conversation which has been tagged with a range of social metadata. Additionally, we will observe how those structures distribute according to a range of social variables.

The present paper builds upon previous work by Culpeper and Gillings (2018) and van Dorst, Gillings and Culpeper (in prep.). In both of these studies, we examined the extent to which different politeness types (solidarity, tentativeness, and deference), and different levels of formality (formal or informal) vary in frequency according to speakers' social background. Two findings from the latter paper, for example, is that the frequency of certain politeness expressions differs depending on the speakers' gender, and depending on the density of the area in which they live. In the present paper, we take a similar methodology, yet focusing this time on impoliteness.

The structures examined in this analysis are a subset of Culpeper's (2010) conventionalised impoliteness formulae. Our selection is driven by the extent to which the structure is searchable via corpus queries, the extent to which it occurs frequently enough to pattern, and the extent to which genuine cases of impoliteness can be found. This reduces the number of structures to just a few, including, for example, the YOU+NP structure, recently examined in Van Olmen, Andersson and Culpeper (2023). All expressions included in the analysis are manually screened via concordance analysis to remove non-genuine cases of impoliteness (e.g., sarcasm), meaning that the analysis allows us to explore exactly how genuine impoliteness varies across speakers.

Our hypothesis, at this stage, is that we would not expect to find much impoliteness in the Spoken BNC2014 (that is, far less than the amount of politeness). The reason for this is

that the types of people who are happy to be recorded for corpus compilation together are likely to get along with each other, and thus we would expect impoliteness rates to be lower. Despite this, preliminary searches suggest that there are cases of genuine impoliteness, and these are the cases that we focus on in the paper.

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

Culpeper, J. (2010). Conventionalized impoliteness formulae. Journal of Pragmatics, 42(12): 3232-3245.

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

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). "The spoken BNC2014: Designing and building a spoken corpus of everyday conversations". International Journal of Corpus Linguistics, 22(3), 319-344.

van Dorst, I., Gillings, M., & Culpeper, J. (in prep.) Pragmatic variation in Britain: a corpus-based study of politeness and social variation. To be submitted to Journal of Pragmatics.

Van Olmen, D., Andersson, M., & Culpeper, J. (2023). Inherent linguistic impoliteness: The case of insultive YOU+NP in Dutch, English and Polish. Journal of Pragmatics, 215: 22-40.

1530 - Talks 7, Syndicate 1

Chair: Mateus De Souza

- 15:30 **Meiqi Li, Ting Jiang** : A Corpus-assisted Discourse Study of Disputants' Intentionality in Chinese Family Mediation

Abstract

In family mediation, disputants present facts and express intentions via languages, but to date their discourse has not received due scrutiny in linguistics. Drawing on the Attitude System, this study used a corpus-assisted approach to investigate disputants' discourse in Chinese family mediation, with an aim to reveal how attitudinal resources are employed to manifest disputants' intentionality. This qualitative and quantitative study is based on a small spoken corpus of six authentic Chinese family mediation sessions with 93,478 Chinese characters. The results indicate that disputants adopt a significant quantity of attitudinal resources to convey their intentionality in Chinese family mediation, particularly at Stage 2 and 3. The language choices of attitudinal resources in disputants' discourse are highly driven by their intentionality, which consists of intention to represent and intention to communicate. The intention to represent attributes disputants' mental state, mental estimation and mental orientation, primarily expressed by the lexico-grammatical level of attitudinal resources; while the intention to communicate represented by Affect, Judgement and Appreciation resources contains more concrete interpersonal meanings depending on the situational context, respectively manifested as touching hearers with emotion, winning hearers by virtue and convincing hearers with reason.

- 15:50 **Liying Zhou** : Comparing functional discourse units in Chinese and English conversational discourse: a cross-linguistic and corpus-based analysis

Abstract

Conversational discourse is frequently examined at a micro-level, focusing on units such as turns and speech acts. Nonetheless, when viewed at a macro level, conversational discourse comprises larger units. As found by Biber et al. (2021) in their empirical examination of Spoken BNC 2014 (Love et al., 2017), conversational discourse is constructed by coherent functional discourse units that can serve a variety of communicative purposes. The current study is based on the framework developed by Egbert et al. (2021), who segmented English conversational discourse into discourse units and classified nine communicative purposes, such as figuring things out, sharing feelings and evaluations, etc. This study used three corpora. The first is the Mandarin Chinese conversation corpus, comprising 420,000 tokens. It was compiled specifically for this study through conversations with 100 native Mandarin speakers. The corpus was annotated in accordance with the framework of Egbert et al. The second corpus used in this research is a British conversation corpus compiled by Fox (forthcoming). The third corpus is the Chinese L1 English Learners corpus. The

current study will first explore the feasibility of applying the taxonomy comprising nine communicative purposes, derived from British English conversations, to conversations in Mandarin Chinese. Additionally, the paper will investigate lexico-grammatical features typical for fulfilling communicative purposes in Chinese conversations using corpus linguistics techniques. Furthermore, this paper will compare and contrast the differences and similarities of discourse units between Chinese conversations and British English conversations to determine how communicative purposes are achieved differently in each language.

- 16:10 **Cristina Lastres-López** : Conditionals in English and Spanish: Exploring differences in speech and writing

Abstract

This paper explores if-conditionals in English and their equivalents in Spanish, introduced by si, delving into the differences in their use in speech and writing. The analysis includes both prototypical conditional constructions in which the protasis indicates the cause and the apodosis expresses the consequence, as in (1), as well as other constructions which encode a wider range of functions in discourse and in which the conditional meaning is weaker (Ford & Thompson, 1986; Ford, 1997; Warchal, 2010; Author, 2020, 2021), as in (2).

(1) Si la inyección económica no viene, lo vamos a pasar mal (CONTRAST-IT cnt_es_mun_dep_006) ‘If the economic injection does not arrive, we are going to have a hard time’

(2) So there’s two different patches there if you see what I mean (ICE-GB:S1A-076 #094:1:B)

Building on prior research on spoken discourse (Author, 2020, 2021), the aim of this presentation is to analyse similarities and differences on the use of these constructions in speech and writing in the two languages examined. The theoretical framework adopted is based on the three metafunctions considered in Systemic Functional Linguistics (Halliday & Matthiessen, 2014), allowing us to distinguish if/si-constructions at the ideational, interpersonal and textual levels. The methodology adopted is corpus-based. Data from spoken discourse are extracted from the conversation subcorpora of the British component of the International Corpus of English (ICE-GB) (Nelson et al., 2002) and the Spanish component of the Integrated Reference Corpora for Spoken Romance Languages (C-ORAL-ROM) (Cresti & Moneglia, 2005). On the other hand, written data are retrieved from the English and Spanish components of CONTRAST-IT (De Cesare, 2018), which comprise data from online newspapers. Preliminary results unveil cross-linguistic differences in English and Spanish, as well as notable dissimilarities in their use in speech and writing.

1530 - Talks 7, Syndicate 2

Chair: Odette Vassallo

- 15:30 **Noelia Ramon, Belén Labrador** : Medals and awards: Providing quality assurance in online promotional discourse

Abstract

In our globalised world, all industries need to promote their products internationally. Companies engaging in global business need to rely on accurate and idiomatic texts to convince potential customers to buy their products. Therefore, a good knowledge of how to operate the expression of high quality in English will greatly benefit non-native professionals involved in promoting their products online. In this paper, we will focus on the subgenre of online promotional texts in the cheese industry, where reference to quality standards is essential. This study investigates the various lexical and phraseological units used by manufacturers and retailers when promoting their products online to identify recurrent patterns that can be taught to non-native speakers of English.

The analysis makes use of a corpus of English Online Cheese Descriptions (OCD). The corpus contains 128,347 words, extracted from different types of British websites dealing with cheeses. OCD was compiled, tagged and explored using software specifically designed for our purposes. The semantic tagging was carried out following the annotation scheme called USAS (UCREL Semantic Analysis System) (Rayson et al., 2004). For this paper, we will analyse instances with the semantic labels of A5.1 Evaluation: good/bad, A5.4 Evaluation: authenticity and S7.3 Competition. All the concordance lines will be investigated from a lexical and phraseological perspective to extract the main linguistic patterns used to express objective quality features of cheeses, including quotations, awards, certifications and medals. These resources contribute to expressing a positive evaluation of the cheeses to be sold (Hunston and Sinclair 2000), thus strengthening the persuasive function of this type of text. The expected results will yield data which can be used to enhance second-language writing for marketing dairy products in English, thus supporting international professionals in multilingual contexts.

References

Hunston, S. and J. Sinclair. 2000. A local grammar of evaluation. in Hunston, S. and Thompson, G. (eds) Evaluation in Text: Authorial Stance and the Construction of Discourse. Oxford: OUP: 74-101.

Rayson, P., Archer, D., Piao, S. and McEnery, T. 2004. The UCREL Semantic Analysis System. In Proceedings of the Workshop Beyond Named Entity Recognition, Semantic Labelling NLP Tasks (LREC 2004), 7–12. Lisbon: European Language Resources Association.

- 15:50 **Cathryn Bennett, Ciara Wigham** : “Can I ask you just to clarify for a minute?”: Teaching pragmatics in IVE pre-service teacher training

Abstract

Intercultural Virtual Exchanges (IVE) have been on the rise in recent years with notable benefits including the integration of cultural and linguistic aspects in second language learning (Dooly & O’Dowd, 2018). The role of pragmatics in teacher-training in online environments has garnered attention as pre-service teacher training programmes aim to integrate IVEs into their curricula. However, challenges may occur due to miscommunication (O’Dowd & Ritter, 2006).

To combat these potential miscommunication episodes, a data-driven pedagogical intervention approach has been recommended (Cunningham & Vyatkina, 2012). While speech acts such as requesting behaviour have been well-researched, there is less evidence in how language for politeness (Alonso-Marks & Bayonas, 2023) has been used to overcome issues in IVEs between second language learners. Therefore, we aim to fill this gap by investigating five miscommunication episodes of pre-service ELF teachers related to politeness strategies to understand how this can be taught in future teacher training programmes.

An IVE with pre-service teachers at French and Dutch universities was carried out in Fall 2022. Pre-service teachers completed three tasks on three different online platforms (Flipgrid, Moodle and Big Blue Button) with the present study focusing on data from the third task of giving feedback on an IVE task for their future classrooms. Video interactions were recorded, transcribed and coded in ELAN to examine participants’ use of linguistics markers for expressing politeness strategies when in communication breakdowns. This led us to our research question of what linguistic markers of politeness are used by ELF IVE pre-service teachers.

Our preliminary results suggest pre-service teachers from the Netherlands express concerns of being overtly ‘direct’ with their French counterparts when providing feedback utilising modal verbs. Our findings outline how these strategies can be integrated in teacher training programmes via exposing learners to linguistic markers of politeness prior to IVE collaborative tasks.

References

Alonso-Marks, E., and Bayonas, M. (2023) Politeness strategies in the greetings of Spanish learners in virtual language learning environments. Borealis – An International Journal of Hispanic Linguistics. 12(2). 387-403. https://doi.org/10.7557/1.12.2.7045.

Cunningham, D.J. (2017) Second language pragmatic appropriateness in telecollaboration: The influence of discourse management and grammaticality. System. 64. 46-57. http://dx.doi.org.10.1016/j.system.2016.12.006.

Cunningham, D.J. (2016) Request Modification in Synchronous Computer-Mediated Communication: The Role of Focused Instruction. The Modern Language Journal. 100(2). 484-507. https://doi.org/10.1111/modl.12332.

Cunningham, D.J., and Vyatkina, N. (2012) Telecollaboration for professional purposes: Towards developing a formal register in the foreign language classroom. The Canadian

Modern Language Review. 68(4). 422-450. https://doi.org/10.3138/cmlr.1279.

Dooly, M., & O'Dowd, R. (2018) Telecollaboration in the foreign language classroom: A review of its origins and its application to language teaching practice. In M. Dooly Owenby & R. O'Dowd (Eds.), In this together: Teachers' experiences with transnational, telecollaborative language learning projects (pp. 11-34). Peter Lang. https://doi.org/10.3726/b14311.

O'Dowd, R., and Ritter, M. (2006) Understanding and Working with 'Failed Communication' in Telecollaborative Exchanges. CALICO Journal. 23(3). 623-642. https://www.jstor.org/stable/24156364.

1530 - Talks 7, Syndicate 3

Chair: Andrew Caines (pre-recorded talks)

- 15:30 **Ljubica Leone** : The representations of freedom in The Sun newspaper between 2019 and 2021: a corpus-based study

Abstract

Existing studies have highlighted the close link between language and society (Fairclough, 1992) and demonstrated the impact of Covid-19 on language (Mahlberg and Brookes, 2021). There are no studies to date that have examined the changing representations and conceptual shifts of freedom in the pre-pandemic and post-pandemic years, which are expected to be affected by government policies on Covid-19.

The present study aims to fill this gap. Specifically, the objective is to examine the conceptual evolution of freedom in the years 2019 and 2021 and to interpret it in light of socio-historical issues derived from the outbreak of the Covid-19 pandemic. The study is a corpus-based investigation undertaken on The Sun Corpus (TS), a corpus including newspaper articles published in the UK in 2019 and 2021. Media including newspapers are influential discourses that shape the public view of particular events, and current issues, and play "an important role in framing how people understand and respond to" contextual happenings (Brookes and Baker, 2021, p. 1; Baker et al. 2013).

The Sun Corpus has been queried with Desktop (offline) corpus analysis tool #LancsBox 6.0 (Brezina et al., 2020). Instances of freedom were retrieved using the KWIC tool in #LancsBox6.0, which allows visualization of concordances. Changing representations of freedom have been examined via collocations extracted with GraphColl.

The analysis reveals that there is a shift in the use of the word freedom which is depicted as an enjoyable experience in 2019 and seen with a negative shade in 2021. These results support the social conceptualization of language and reveal aspects that are of particular

concern in Critical Discourse Analysis (Fairclough, 1992) aiming to examine how socio-historical aspects frame the linguistic representations of social issues like the Covid-19 disease.

References

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

Brezina, V., Weill-Tessier, P. and McEnery, A., 2020. #LancsBox v.6.[software package]. http://corpora.lancs.ac.uk/lancsbox

Brookes, G. and Baker, P., 2021. Obesity in the news: Language and representation in the press. Cambridge: Cambridge University Press.

Fairclough, N., 1992. Discourse and social change. Cambridge: Polity Press.

Mahlberg, M. and Brookes, G., 2021. Language and Covid-19. Corpus linguistics and the social reality of the pandemic. International Journal of Corpus Linguistics, vol.26, no. 4, pp. 441–443.

- 15:50 **Eniko Csomay** : Teaching styles and the 'mixed' category

Abstract

Teaching styles in educational studies have been associated with the needs, beliefs, and behaviors that teachers display in the classroom. These studies use perceptual measures to illustrate how teachers present information, interact with students, and supervise coursework but provide no linguistic analysis. Corpus linguistic studies of university classroom discourse reflect two approaches to the analysis: researchers take individual or a select group of linguistic features to explore their functional variants in academic lectures, or, relying on emerging, co-occurring linguistic patterns, they provide comprehensive linguistic characterizations. However, classifications of teaching styles based on linguistic measures are still lacking.

This study reports on the findings of a corpus-driven investigation of linguistic patterns in university class sessions that are associated with teaching styles. The parameters of a situational analysis are followed by a multivariate statistical analysis and a K-means cluster analysis. Taking the transcripts of over 194 class sessions in North American classrooms, the co-occurrence patterns of over 200 linguistic features are tracked. The results of the factor analysis are interpreted functionally underlying the dimensions of linguistic variation which then serve as the basis for the cluster analysis. Five clusters are associated with teaching styles:

1. `Topic-focused instruction with abstract and quantity information`

2. Narrative instruction with explicit stance
3. Expository instruction with persuasion
4. Mixed styles of instruction
5. Context-oriented instruction lacking explicit stance.

While teaching styles could be identified via the functions of linguistic correlates in four categories, it was more difficult to interpret the category called “mixed styles of instruction”. Issues with classifications into discrete categories when it comes to university teaching or teaching in general (whether based on perceptual measures in educational research or taking a corpus-driven linguistic approach to the analysis) will also be discussed.

- 16:10 **Jingwen Ou** : Introducing data-driven learning into Chinese higher education EAP writing instructional settings

Abstract

Writing for academic purposes in a second or foreign language is one of the most important yet the most demanding skills for non-native speakers. Traditionally, the EAP writing instruction at tertiary level encompasses the teaching of the disciplinary writing conventions, the rhetorical functions, and specific linguistic features (Kobayashi & Rinnert, 2012). However, one of the main challenges in English academic writing for L2 students at tertiary level can still be found in the mastery of academic discourse (Lin & Morrison, 2021). The past two decades have witnessed a rising popularity of data-driven learning (DDL) approach in the field of EAP writing teaching (Chen and Flowerdew, 2018). Such a combination has significantly transformed traditional pedagogy and triggered global research on corpus use (Crosthwaite et al, 2021).

Given the uprising popularity of DDL in EAP classrooms, my research aims to investigate Chinese university students’ use of corpus tools in the higher education setting, specifically in EAP writing contexts, with three main foci: 1) influence of corpus tools on learning behaviours, 2) influence of corpus tools on students’ academic writing performance outcomes, and 3) students’ perceptions and potential perceptional changes towards the use of such tools. It adopts a three-part workshop training design including the training on the use of three corpus tools: CQPWeb, Sketch Engine and LancsBox X 4.0.0.. Three training workshops focus on the instruction of selected corpus functions and most commonly used rhetorical functions in different essay sections in research articles.

This study adopts an experimental design, where twenty participants will be assigned randomly to an Intervention Group (N=10) and a Non-intervention Group (N=10). A pre-test and an online survey are administered before intervention to gather information on participant profile, original attitudes towards DDL and academic writing proficiency. Each training workshop is followed by collections of essay sections and a focus group interview.

Insights from the preliminary public involvement and a pilot test with Chinese postgraduate students suggested students' notably positive attitudes towards the use of corpus tools for English learning, particularly for improving their collocational accuracy, contextual interpretation of the searched words, and linking rhetorical functions to their academic communicative purposes. Students also reported the convenience of visualization functions in LancsBox X and the clear categorization of collocations in Sketch Engine. However some challenges were also identified during the training , including 1) insufficient time to complete essay revision, 2) struggle with technical set-up, 3) unfamiliarity with DDL approach, 4) difficulties with advanced corpus functions, such as reading different Association Measures in CQPWeb and generating collocation graph with LancsBox X, 5) low efficiency in revision due to a lack of feedback on their drafts, and 6) a lack of cognitive preparedness for connecting rhetorical functions with essays. Such struggles indicated the direction for customizing solutions in the main study and offered insights to EAP writing practitioners for incorporating DDL tools with carefully scaffolded tasks to enhance students' understanding and use of rhetorical functions.

References

Chen, M., & Flowerdew, J. (2018). A critical review of research and practice in data-driven learning (DDL) in the academic writing classroom. International Journal of Corpus Linguistics, 23(3), 335–369. https://doi.org/10.1075/ijcl.16130.che

Crosthwaite, P. & Sanhueza, A. G. & Schweinberger, M. (2021). Training disciplinary genre awareness through blended learning: An exploration into EAP students' perceptions of online annotation of genres across disciplines, Journal of English for Academic Purposes, 53. https://doi.org/10.1016/j.jeap.2021.101021.

Kobayashi, H. & Rinnert, C. (2012). Chapter 5. Understanding L2 writing development from a multicompetence perspective: Dynamic repertoires of knowledge and text construction. In R. Manchón (Ed.), L2 Writing Development: Multiple Perspectives (pp. 101-134). Berlin, Boston: De Gruyter Mouton. https://doi.org/10.1515/9781934078303.101

Lin, Linda H.F., & Morrison, B. (2021). Challenges in academic writing: Perspectives of Engineering faculty and L2 postgraduate research students. English for Specific Purposes, 63, 59-70.

1630 - Conference Close