Corpus Linguistics: Past, Present, Potential

Presentation Abstracts

13th May 2022

University of Leeds

Isabelle Clarke, Lancaster University

Approaching Discourse in Corpora using Multiple Correspondence Analysis

In this talk, I will introduce a new approach to the analysis of keywords (Clarke et al. 2021) and demonstrate how this approach can be used to examine variation in discourses over time (Clarke et al. forthcoming). Keywords can offer analytical signposts to discourses in large corpora, yet keywords do not map straightforwardly onto discourses. One major challenge with keyword analyses is aggregation – the keywords may all be associated with discourses present in the corpus but disaggregating the discourses in the keywords is a matter for the analyst. To make analysis more manageable, keywords are often grouped manually into semantic or thematic categories. Different approaches have been taken to achieve this; however, all approaches have limitations, and the creation of meaningful categories and the assignment of keywords to those categories often involves some element of compromise, especially as keywords may contribute to more than one discourse. Determining where this happens is, again, a matter for the analyst. Moreover, the processes for choices made in categorisation may not be explicitly documented or implied rather than conforming to a set method. Another persistent issue with keyword studies is their focus on their presence rather than absence, yet absence can be as meaningful as presence in discourse analysis (Schroeter & Taylor 2018) and patterns of presence and absence across a corpus may meaningfully interact (Partington 2014).

As a result, I developed a new approach with Tony McEnery and Gavin Brookes, called 'Keyword Co-occurrence Analysis (KCA)', which draws on Multiple Correspondence Analysis to group keywords statistically based on how they co-occur with each other in the texts of the corpus. My talk will show how KCA overcomes many of the issues in traditional keyword analyses, including absence, and has proven to be effective for providing a more nuanced account of keywords that is sensitive to the various senses and discourses that a single keyword can exhibit. Moreover, I will demonstrate the potential of the approach for investigating discourses over time.

Clarke, I., McEnery, T & Brookes, G. (2022). Keywords through time: Tracking changes in press discourses of Islam. *International Journal of Corpus Linguistics, forthcoming.*

Clarke, I., McEnery, T., & Brookes, G. (2021). Multiple Correspondence Analysis, Newspaper Discourse and Subregister: A Case Study of Discourses of Islam in the British Press. *Register Studies 3(1):* 144—171.

Partington, A. (2014). Mind the gaps: The role of corpus linguistics in research absences. *International Journal of Corpus Linguistics 19(1):* 118—146.

Schroeter, M. & Taylor, C. (2018). *Exploring Silence and Absence in Discourse: Empirical Approaches.* London: Palgrave Macmillan.

David Wright, Nottingham Trent University

Corpora and the evolution of forensic linguistics

This talk examines the role that corpora and the use of corpus linguistic approaches have played in the expansion and development of forensic linguistics since the 1990s. As with many disciplines of language analysis, forensic linguistics has benefitted from access to larger datasets and the effective combination of quantitative and qualitative analyses of these datasets. This is especially true for areas of the field for which large amounts of data are readily available. However, many types of data that forensic linguists are often interested in can be difficult to access due to their sensitive and personal content or their general rarity in the public domain. This means that researchers committed to using corpus linguistics in forensic contexts are faced with challenges that are not as profound in other corpus-assisted disciplines. This talk will give an overview of how this apparent challenge has in fact given rise to developments that have advanced the study of forensic linguistics, including the use existing general reference corpora and publicly available data, efforts underway to make new forensically relevant corpora available to researchers, and, most notably, the emergence of new directions for the field that widen the scope of 'forensic linguistics'. In relation to the last point, a case study is presented in which corpus-assisted discourse analysis is used to examine how the proposal and passage of new laws has been reported in the British national press in the last twenty years. This case study is offered as an example of how forensic linguistics may be considered a socio-legal discipline, situating the study of language and law within its broad social, political and cultural context.

Eric Atwell, University of Leeds

AI4AI: Artificial Intelligence for Arabic and Islamic Corpus Linguistics

What is challenging about Arabic and Islamic text? Corpus concordancers and taggers do not adapt easily to Arabic script and morphology. Modern Standard Arabic is a de-facto world-wide common written form, but a bewildering range of spoken dialects are used in social media and informal publications. Arabic and Islam are symbiotes: Arabic is the main language of Islam, and Islamic words and concepts permeate Arabic text. At Leeds University, we are applying Artificial Intelligence to these and other challenges of Arabic and Islamic corpus linguistics (Atwell 2019). Recent research includes: collecting and annotating Arabic Hadith texts (Altammami et al 2019a,b, 2020a,b, Tarmom et al 2019, 2020a,b,c), and using ontologies, machine learning and deep learning to model semantics of Arabic Quran and Hadith texts (Alsaleh et al 2021, Alshammeri et al 2022, Alshammeri et al 2020, 2021a,b,c, Altammami 2020c,2022, Liu et al 2019).

Alsaleh AN, Atwell E, Altahhan A. 2021. Quranic Verses Semantic Relatedness Using AraBERT. The Sixth Arabic Natural Language Processing Workshop (WANLP 2021) Proceedings of the Sixth Arabic Natural Language Processing Workshop , pp. 185-190

Alshammari I, Atwell E, Alsalka MA. 2022. Automatic Mapping of Quranic Ontologies Using RML and Cellfie Plugin. NLDB 2022 : The 27th International Conference on Natural Language & Information Systems, Springer

Alshammeri M, Atwell E, Alsalka MA. 2020. Quranic Topic Modelling Using Paragraph Vectors. 2020 Intelligent Systems Conference (IntelliSys) Advances in Intelligent Systems and Computing Springer Verlag 1251, pp. 218-230

Alshammeri M, Atwell E, Alsalka MA. 2021a. A Siamese Transformer-based Architecture for Detecting Semantic Similarity in the Quran. 9th International Conference on Islamic Applications in Computer Science and Technologies (IMAN 2021)

Alshammeri M, Atwell E, Alsalka MA. 2021b. Classifying Verses of the Quran using Doc2vec. The 18th International Conference on Natural Language Processing (ICON2021) ACL Anthology ACL

Alshammeri M, Atwell E, Alsalka MA. 2021c. Detecting Semantic-based Similarity Between Verses of The Quran with Doc2vec. Fifth International Conference On AI In Computational Linguistics Procedia Computer Science Elsevier 189, pp. 351-358

Altammami S, Atwell E, Alsalka A. 2019a. Text Segmentation Using N-grams to Annotate Hadith Corpus. The 3rd Workshop on Arabic Corpus Linguistics Proceedings of the 3rd Workshop on Arabic Corpus Linguistics Association for Computational Linguistics, pp. 31-39

Altammami S, Atwell E, Alsalka A. 2019b. The Arabic–English Parallel Corpus of Authentic Hadith. International Conference on Islamic Applications in Computer Science and Technologies – Proc IMAN'2019

Altammami S, Atwell E, Alsalka A. 2020a. Constructing a Bilingual Hadith Corpus Using a Segmentation Tool. LREC 2020 Proceedings of The 12th Language Resources and Evaluation Conference The European Language Resources Association (ELRA), pp. 3390-3398

Altammami S, Atwell E, Alsalka A. 2020b. The Arabic–English Parallel Corpus of Authentic Hadith. International Journal on Islamic Applications in Computer Science And Technology 8(2), pp. 1-10

Altammami S, Atwell E, Alsalka A. 2020c. Towards a Joint Ontology of Quran and Hadith. International Journal on Islamic Applications in Computer Science And Technology.

Altammami S, Atwell E. 2022. Challenging the Transformer-based models with a Classical Arabic dataset: Quran and Hadith. Proc LREC'2022

Atwell E. 2019. Using the Web to model Modern and Qurʾanic Arabic. In: McEnery T; Hardie A; Younis N (eds.) Arabic Corpus Linguistics. Edinburgh, UK: Edinburgh University Press, pp. 100-119

Liu Z, Yang L, Atwell E. 2019. The Semantic Annotation of the Quran Corpus Based on Hierarchical Network of Concepts Theory. International Conference on Asian Language Processing (IALP) IEEE, pp. 318-321

Tarmom T, Atwell E, Alsalka M. 2019. Non-authentic Hadith Corpus: Design and Methodology. International Conference on Islamic Applications in Computer Science and Technologies (IMAN 2019) Proceedings of IMAN 2019

Tarmom T, Atwell E, Alsalka MA. 2020a. Automatic Hadith Segmentation using PPM Compression. ICON-2020 Proceedings of the 17th International Conference on Natural Language Processing (ICON-2020) Association for Computational Linguistics (ACL)

Tarmom T, Atwell E, Alsalka MA. 2020b. Non-authentic Hadith Corpus: Design and Methodology. International Journal on Islamic Applications in Computer Science And Technology. 8(3), pp. 13-19

Tarmom T, Teahan W, Atwell E, Alsalka MA. 2020c. Compression versus traditional machine learning classifiers to detect code-switching in varieties and dialects: Arabic as a case study. Natural Language Engineering. 26(6), pp. 663-676

Richard Badger, University of Leeds

Agency and action: a study of the UK Government's press conferences on Covid-19

The Covid-19 pandemic has highlighted the importance of effective communication about health care issues, with significant differences in how diverse groups have responded to government information, but we have little information about how the substance of those communications impacts on their transparency and trustworthiness for the public.

The Organisation for Economic Co-operation and Development (OECD) identifies the centrality of transparency and trustworthiness in health related communication (OECD, 2020). However, we have little information of what linguistic and multimodal features make communications transparent and trustworthy. This study investigates whether a linguistic/multimodal analysis might enable us to identify what these features might be and relates them to public perceptions for the levels of trustworthiness and transparency of the press conferences.

This paper reports on the linguistic part of the study. We constructed a corpus of UK Government ministerial speeches that were a part of the press conferences about the Coronavirus from 3 March 2020 to 5 April 2021. The corpus consists of 128 documents and totals just over 150, 000 words.

The study investigated transparency through readability formulae (Crossley et al., 2017; Munley et al., 2018) and the clause as representation, particularly process types, within

functional linguistics (Halliday & Matthiessen, 2014). Trustworthiness was also investigated using systemic functional linguistics focusing on mood and pronoun use.

The aim is to use this study as the basis for a bid to Nuffield. The current research team comprises Jimmy Choo (Healthcare), Matt Homer (Education), Elisabetta Adami (Languages, Cultures & Societies) and Richard Badger (Education). The original team for this project included Martin Thomas and we would like to acknowledge the contribution he made to the project before his death.

Crossley, S. A., Skalicky, S., Dascalu, M., McNamara, D. S., & Kyle, K. (2017). Predicting Text Comprehension, Processing, and Familiarity in Adult Readers: New Approaches to Readability Formulas. *Discourse Processes*, *54*(5-6), 340-359. https://doi.org/10.1080/0163853X.2017.1296264

Halliday, M. A. K., & Matthiessen, C. (2014). *Halliday's Introduction to functional grammar* (Fourth ed.). Routledge.

Munley, B., Buser, A. T., Gaudreau, S., Breault, J. L., & Bazzano, L. A. (2018). An Analysis of Informed Consent Form Readability of Oncology Research Protocols. *Journal of Empirical Research on Human Research Ethics*, *13*(4), 363-367. https://doi.org/10.1177/1556264618795057

OECD. (2020). Transparency, communication and trust: The role of public communication in responding to the wave of disinformation about the new Coronavirus. *OECD Policy Responses to Coronavirus (COVID-19)*. https://doi.org/doi.org/10.1787/bef7ad6e-en . (OECD Policy Responses to Coronavirus (COVID-19))

Yen Dang, University of Leeds

How can corpus linguistics help to tackle the lexical challenge of academic listening?

To achieve academic success in English-medium university programs, students need to understand not only their reading materials but also lectures, seminars, labs, and tutorials. Yet comprehending academic spoken English is challenging for many second language (L2) learners, and insufficient vocabulary knowledge is frequently cited as a major reason for this difficulty.
Despite the significant role of vocabulary in comprehension, vocabulary in academic spoken English is an under-explored area of vocabulary studies. In the talk, I will share the findings of some studies of mine which used corpus linguistics to develop word lists and identify resources for L2 learners to learn academic spoken vocabulary.

Alice Deignan, University of Leeds, and Duygu Candarli, Dundee University

Using corpora to support UK school students

In early phases of corpus linguistics, we developed important insights into the nature of language, such as the centrality of lexis and collocation, the interaction between grammar and lexis, and the vast variation in language use across registers. More recently, many corpus linguists have started looking outwards, at how their discipline interacts with others, and how it can contribute to society. Our research is an interaction between linguistics and education, and creates knowledge that can be used to support school students in the UK.

We present a project (The linguistic challenge of the transition from primary school to secondary school, funded by ESRC, grant number ES/R006687/1) that researches the linguistic challenges facing school students when they move from primary to secondary school, based on data gathered in the north of England. We have worked with five secondary schools and eight primary schools to build corpora that aim to represent the

language encountered by students in Years 5-6, and in Years 7-8, in academic contexts. We are in the process of developing a detailed description at the lexical, grammatical and discourse levels of the spoken and written language of Key Stage 3, contrasted with Key Stage 2 and with language outside the school. Our database allows for comparisons at the broadest level between late KS2 and early KS3, between separate year groups, and between individual subjects at different levels, and we are working on a number of studies.

This talk will focus firstly on changes in the written materials that are encountered by school students when they start secondary school, and secondly, on changes in the vocabulary and structures found in the materials used for teaching.

Fiona Douglas, University of Leeds

The divisive language of union: from indyref to Brexit

Indyref and Brexit form the bookends of a series of seismic political events that took place from 2014 to 2020 and shook the UK political landscape to its core. In 2014, Scotland voted on whether to remain part of the United Kingdom, narrowly voting by 55% to 45% to stay. Two years on, by an even narrower margin (52% to 48%, though voting patterns differed around the country), the UK voted to leave the European Union − and Brexit became a reality. The arguments have been characterised by increasingly divisive language, which many have decried as damaging. Centrifugal politics are rife, and the ever-widening spirals of disunity look likely to persist within the UK and beyond. What began as friction, quickly led to fragmentation, and may end in fracture.

Based on three specialist corpora totalling over 143 million words and comprising a wide range of different text types including newspaper articles, speeches, televised and parliamentary debates, party websites, Twitter postings and observations from political pundits, this paper offers a diachronic perspective on the political, media and public discourses surrounding two unions − the United Kingdom and the European Union. It explores what identities were invoked during and after these referendum campaigns (national, supranational, social, ethnic or racial?), how these were perceived by media and electorates, and ultimately the relationship between politics and the power of language to unite or divide.

Mel Evans, University of Leeds

Unprecedented Communities, Discourses, and Evolving Identities of #LongCovid

This paper argues for the value of diachronic approaches to discourse, and the capacity of corpus linguistics to highlight continuities and differences over time in current, as much as historical, matters of language and society. I present some preliminary findings of an interdisciplinary study that investigates discourses of long covid, aiming to produce best practice communication guidelines for patients and healthcare practitioners managing the condition. Long Covid was identified through patient activism, enabled via social media networks, such as Facebook and Twitter (Perego et al. 2020). I use corpus linguistic methods to investigate the ways in which language has been used to conceptualise, reify and evaluate Long Covid as a condition, using a longitudinal Twitter corpus.

The corpus includes tweets with the hashtag #LongCovid from three time points during the early stages of the covid-19 pandemic: July 2020, October 2020 and March 2021. Drawing on the concept of ambient communities (Zappavigna 2011), I propose that sociolinguistic perspectives on social media offer a nuanced understanding of nascent discourses around Long Covid. I focus on naming strategies around the condition and entities affiliated with it (e.g. patients, government, professionals), tracking the trajectory of labels across the dataset. Through the examination of collocates and lexical bundles, I evaluate how facets of power (local, institutional) inform the uptake of an emergent, and ideologically significant, terminology around Long Covid, and more generally reflect on the value of a diachronic perspective on discourse using both qualitative and quantitative corpus linguistic techniques.

Perego, E., Callard, F., Stras, L., Melville-Jóhannesson, B., Pope, R. and Alwan, N.A. 2020. Why the Patient-Made Term 'Long Covid' is needed. *Wellcome Open Research*. **5**, p.224.

Zappavigna, M. 2011. Ambient affiliation: A linguistic perspective on Twitter. *New Media & Society*. **13**(5), pp.788–806.

Alison May, Mashael AlAmr, Yan Chen, Maram AlRabie and Min (Kayla) Wang, University of Leeds

Corpus-based research in Forensic Linguistics: data, methods, and knowledge creation.

Forensic linguistic research has an aspirational aim to use language analysis to 'improve the delivery of justice' (Grant and MacLeod 2020: 180). Rigorous linguistic analysis is at the heart of this endeavour, as, when language becomes evidential, forensic linguists need to ensure that their opinions are based on secure knowledge. In order to do that, we need to know much more about how language really works. This pecha kucha style presentation aims to show the variety and scope of corpus-based forensic linguistic research through the work of five researchers. Beginning with our data (e.g. Arabic Twitter, the Derek Chauvin murder trial, a corpus of authors writing in different genres, a corpus of Chinese criminal trials) we raise important questions in relation to current research priorities in the fields of courtroom discourse, police interviewing and authorship attribution. Moving on to methodologies, using show and tell, we showcase a range of computational tools and methods for exploiting data searching and analysis, including corpus approaches to video trial data, searching for question types, and identifying collocational frameworks. In the final part of our presentation we show some of the results of our research that contribute to knowledge creation in the field of forensic linguistics, specifically, but language analysis more generally. Using the pecha kucha style, we ultimately aim to give a brief flavour but a broad insight into our work that might stimulate discussion.

Serge Sharoff, University of Leeds

Does Size Matter? Statistics from Large Corpora

The number of available corpora grows, so that we can choose working with very large corpora (enTenTen20 has 36 billion words). However, if we want to explore a specialised topic, for example, COVID19 misinformation, we are likely to work with only a small sample. The other side is that enTenTen20 has no metadata apart from the web page URLs, while a

smaller corpus, such as the BNC, can be annotated with respect to its genres, so that it is possible to compare its subsets, such as fiction vs newspapers. In my presentation I will address the issue of variation in corpora with respect to topics and genres as well as an estimation of their match to the purposes of corpus use, for example, for teaching foreign languages. Quite predictably larger corpora can produce better statistically significant models, but sometimes at the expense of mismatch with the purposes of their use.

James Wilson, University of Leeds

"The corpus class", "from ad-hoc corpora to personalised keyword lists" and "students as linguistic detectives": A Look at How IntelliText Has Been Integrated in Russian Language Teaching at Leeds

This presentation focuses on the use of the IntelliText interface (http://corpus.leeds.ac.uk/it) in Russian language and linguistics modules at Leeds from its launch in 2011 to the present day. Since its inception, IntelliText has reshaped the way in which Russian language at Leeds is taught and learned, in particular enhancing possibilities for personalised and autonomous language learning, and forms the basis of a newly introduced linguistics module, The Structures of Russian, exploring grammatical variation in Russian. Students have benefited from reference materials and "hands-off" exercises informed by corpus data and, at the more advanced levels, from direct engagement with corpora in the classroom. IntelliText not only helps students improve their language proficiency, but regular "hands-on" work with IntelliText hones students' linguistic awareness and allows learners to unravel language problems, some of which are not covered in standard language manuals or grammar books, without the input of a tutor, thus equipping them with autonomous learning skills that are essential beyond the degree. I discuss the use of IntelliText in Russian language teaching at Leeds for different language learning purposes. Examples include the use of IntelliText to support traditional approaches to grammar teaching, consolidating and deepening knowledge of grammatical structures with a focus on successfully integrating them in students' writing, to facilitate vocabulary acquisition and to raise students' awareness of grammatical variation in Russian and the relationship between prescriptive rules and actual language use.