Workshop 2: Diversity and Innovation in Concordance Organisation and Interpretation

ABSTRACTS

Natalie Finlayson & Michaela Mahlberg (University of Birmingham) ***Concordancing in the 21st century: A brief review of current practices***

At the beginning of the century, Teubert (2001: 125–126) warned of a downside to the rapid expansion of corpus linguistics, noting that an inward-looking focus on corpus construction and data standardisation may come at the expense of furthering "the original gain that the analysis of corpora may contribute to our knowledge of language." Not unrelatedly, Sinclair (2003) pointed to a need for reliable methodological procedures in anticipation of increasing amounts of concordancing work being carried out computationally. How such a framework might look in practice is currently unknown, but its development represents a crucial step towards moving the discipline forward in a time of renewed growth and technological change.

In this paper, we ask what still needs to be done to bring a level of systematicity to concordance reading that aligns with the flexibility and popularity of the approach and the technical innovations of the present day. As a starting point, we illustrate the variety of ways in which analysts select, organise, and interpret concordance data with examples from literature in four disciplines that bring different motivations and assumptions to the process: lexicography, data-driven learning, corpus-assisted discourse analysis, and literary stylistics. Our overview builds on a small body of work (e.g., Sinclair, 2003; Anthony, 2018; Gillings & Mautner, 2023; Hanks, 2013; Hoey, 2005; Hunston & Francis, 2000; Mahlberg, 2005) that lays the foundations for the development of a structured concordancing methodology based on principled choices about *what* information analysts want to see, *how* they want to see it, and *how* they will make sense of it. By mapping analysts' decisions and considering how their concordancing methods are driven by practical and theoretical contexts, our review not only enhances our understanding of trends that characterise disciplinary practices but also offers insights into three fundamental strategies that underpin concordancing work more broadly. Most strategies can be described as a means of creating a *subset* of data to be analysed, *ordering* concordances so that patterns can be revealed more easily, or *grouping* concordance lines in preparation for interpretation with reference to linguistic and other frameworks.

We envisage that discussions in today's workshop will build on these beginnings, paving the way for systematic, transparent, and much-needed theoretical, methodological, and technical innovation in each of the three areas identified.

References

Anthony, L. (2018). Visualization in corpus-based discourse studies. In C. Taylor & A. Marchi (Eds.), *Corpus approaches to discourse: A critical review* (pp. 197–224). Routledge.

Gillings, M., & Mautner, G. (2023). Concordancing for CADS: Practical challenges and theoretical implications. *International Journal of Corpus Linguistics.*

Hanks, P. (2013). *Lexical analysis: Norms and exploitations*. MIT Press.

Hoey, M. (2005). *Lexical priming: A new theory of words and language*. Routledge.

Hunston, S. & Francis, G. (2000). *Pattern grammar: A corpus-driven approach to the lexical grammar of English*. John Benjamins.

Mahlberg, M. (2005). English general nouns: A corpus theoretical approach. John Benjamins.

Sinclair, J. (2003). *Reading concordances*. Pearson.

Teubert, W. (2001). Corpus linguistics and lexicography. *International Journal of Corpus Linguistics, 6*(1), 125–153.

Mathew Gillings (Vienna University of Economics and Business) ***Concordance analysis in CADS: Does "expanding the line" really work?***

Located at the intersection between quantitative and qualitative approaches to textual analysis, concordance analysis is one of the main techniques within a corpus linguist's toolkit. However, despite a growing body of work critically exploring previously unquestioned mainstays of corpus methods (Mautner, 2015; Taylor & Marchi, 2018; Gillings et al., 2023), it is rare to see this applied to concordance analysis specifically. One recent example of such work is Gillings & Mautner (2023), which explored the range of different issues that researchers may encounter when interpreting concordances within a corpus-assisted discourse analysis (CADS) framework. Drawing on an almost 20-million-word corpus of every article and book review published in *Administrative Science Quarterly* from 1956–2018 (Mautner & Learmonth, 2020), the paper identified eight key issues in concordance line interpretation: noise in the corpus, non-standard syntax, unclear referring expressions, unclear quotation source attribution, technical terms and jargon, acronyms and initialisms, unspecific co-text, and finally lines that are unrelated to the research question. Around one quarter of all lines analysed were uninterpretable; a number that is perhaps relieving or surprising, depending on what exactly one uses concordance analysis for.

For those who use concordance analysis to aid in (critical) discourse analyses specifically, this is likely to be surprising, and a problem. After all, the key remit is to get a sense of the range of different views and representations in a corpus, regardless of whether they are frequent or not. There are few solutions for what to do with uninterpretable concordance lines. Weisser (2016) suggests removing such lines from the analysis (provided such decisions are properly documented), whilst Collins (2019) suggests either extending the span of the co-text or revisiting the full text. These solutions are centred around either increasing the amount of co-text that is viewed or being openly transparent about removing them. Collins' advice to "expand the concordance line" is commonly cited in corpus linguistics literature.

This talk explores the extent to which this advice works in practice. Does "expanding the concordance line" really help? Returning to the uninterpretable lines identified by Gillings and Mautner (2023), I examine what additional steps are necessary to make them interpretable focussing on which of the eight key issues are potentially salvageable and which continue to be a problem. Preliminary analyses suggest that interpretability issues due to unclear referring expressions, unclear quotation source attribution, and unspecific co-text can often be solved by expanding the concordance line. Other lines, however, require further digging either elsewhere in the corpus, or from outside of it. The talk concludes with some thoughts on how developers of concordancing systems may aid (or indeed fix) these issues.

References

Collins, L. (2019). Corpus linguistics for online communication: A guide for research. Routledge.

Gillings, M., & Mautner, G. (2023). Concordancing for CADS: Practical challenges and theoretical implications. *International Journal of Corpus Linguistics.* https://doi.org/10.1075/ijcl.21168.gil

Gillings, M., Mautner, G., & Baker, P. (2023). *Corpus-assisted discourse studies.* Cambridge University Press.

Mautner, G. (2015). Checks and balances: How corpus linguistics can contribute to CDA. In R. Wodak & M. Meyer (Eds.), *Methods of critical discourse studies* (3rd ed.) (pp. 154–179). SAGE.

Mautner, G., & Learmonth, M. (2020). From *administrator* to *CEO*: Exploring changing representations of hierarchy and prestige in a diachronic corpus of academic management writing. *Discourse & Communication, 14*(3): 273–293.

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

Weisser, M. (2016). *Practical corpus linguistics: An introduction to corpus-based language analysis.* Blackwell.

Susan Hunston (University of Birmingham) & Xin Susie Sui (Capital Normal University)
Modelling the output from concordance lines

Although concordancing has a very long history, it is the Key Word in Context format of concordance lines that is associated with Corpus Linguistics. Concordance lines tend to be somewhat marginalised in Corpus Linguistics research, with their significance limited to (a) finding information of importance to lexicography, (b) checking the results of quantitative studies, and (c) finding examples of phenomena identified by other means. However, the output from studies of concordance lines has had a considerable impact on models of language that have either emerged from or been substantially influenced by the study of corpora.

The starting point for this paper is Sinclair's work in the 1980s that developed concepts such as the Unit of Meaning and the Idiom Principle (Sinclair, 1991: 2004). This work focused on lexis and grammar as a single system, on the unity of form and meaning, and on the location of meaning in the phrase rather than in the individual word. Sinclair demonstrated his approach in a series of specific word-studies (2003, 2004), and the Collins COBUILD series of dictionaries and grammars provided detailed descriptions of English using the same principles. The work was extended and given a further theoretical perspective, by, for example, Lexical Priming (Hoey, 2003), Local Grammar (Barnbrook, 2002; Cheng & Ching, 2016), and Corpus Pattern Analysis (Hanks, 2013). The scrutiny of concordance lines by individuals was the key methodology used in each case.

Sinclair, however, was not alone in recognising the interconnectedness of form and meaning, lexis and grammar. The concept of the Construction (Goldberg, 1995; Hoffman & Trousdale, 2013) developed independently of the Unit of Meaning, but is very similar to it, in particular in its rejection of the lexis-grammar distinction and its identification of meaning with form. Many of the examples of Units of Meaning discussed in the literature could be described as Constructions, and vice versa. The FrameNet project (Fillmore et al., 2003), with its mapping of meaning to form, shares much with the notion of Local Grammar, even though, again, they developed independently and largely unaware of each other. In consequence, there are multiple approaches that are similar but not identical, taking different theoretical standpoints and focusing on distinct but overlapping language phenomena. They all have a starting point in the scrutiny of large amounts of naturally-occurring language, with concordance lines at the heart of this.

This paper tries to make sense of this muddle of terminology and proposes an approach to thinking about four concepts – Units of Meaning, Local Grammar, FrameNet and Construction Grammar – that clarify what they share and how they differ. A series of oppositions is used to make these comparisons: mental focus vs output focus; form-to-meaning vs meaning-to-form; notion focus vs function focus; general vs partial theory; specific vs non-specific context. The result is a step-wise model that traces a progression of thinking from observation of concordance lines to contextualised theories of language.

Acknowledgement: This study is partially supported by the MOE Project of Humanities and Social Sciences (Project No. 19YJC740069) and the China Scholarship Council (File number: 202307300026).

References

Barnbrook G. (2002). *Defining language: A local grammar of definition sentences*. Benjamins.

Cheng, W., & Ching, T. (2016). 'Not a guarantee of future performance': the local grammar of disclaimers. *Applied Linguistics*, *39*(3), 263–301.

Fillmore C., Johnson C., & Petruck, M. (2003). Background to FrameNet. *International Journal of Lexicography*, *16*(3): 235–250.

Goldberg, A. (1995). *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Hanks, P. (2013). *Lexical analysis: Norms and exploitations.* MIT Press.

Hoey, M. (2005). *Lexical priming: A new theory of words and language*. Routledge.

Hoffman T., & Trousdale, G. (2013). *The Oxford handbook of construction grammar*. Oxford University Press.
Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.
Sinclair, J. (2003). *Reading concordances*. Longman.
Sinclair, J. (2004). *Trust the Text: Language, corpus, and discourse*. Routledge.

Laurence Anthony (Waseda University) ***A sentence embedding approach to concordance searching and sorting***

Concordancing has long been a cornerstone of corpus linguistics research, providing scholars with a powerful method to explore lexical and grammatical patterns in target corpora. It is also one of the most common approaches introduced to learners in a data-driven learning (DDL) classroom. Despite the strengths of the approach, it also suffers from two major limitations. Firstly, concordance searching requires the use of single or multi-word queries that are often fixed in nature and can quickly increase in complexity depending on the aim. For example, to account for possible variations in usage, these queries usually require the use of alternative options or the inclusion of in-word or between-word wildcards. If the researcher, teacher, or learner hopes to capture subtle variations in usage in the corpus (e.g., spelling differences between UK and US speakers, idiomatic expression with synonym variations, semantically equivalent words or phrases), these differences have to be recognized from the outset and accounted for in the query.

A second limitation of concordancing relates to the sorting of results. Typically, results are sorted alphabetically on the center (node) word, or words to the left or right of the node word. This ordering leads researchers, teachers, and learners to have to scan through all results to find relevant, salient patterns of usage. Recently, we have seen innovations such as KWIC patterns (Anthony, 2018, 2022) that calculate the frequency of occurrence of concordance result patterns and order the results accordingly. However, even here, if the query generates many thousands of hits for a particular pattern, there is still a need to sort these results in some meaningful way before they can be interpreted.

Over the past year, much attention has begun to focus on the potential impact of Artificial Intelligence (AI) on corpus research. In this paper, I introduce an innovative approach to concordance querying and sorting that integrates traditional concordance methods with transformer-based sentence (or sentence fragment) embeddings. Using sentence embeddings, I show how concordance search queries can be greatly simplified and also allow for more nuanced and context-aware analysis of linguistic phenomena than previously possible. In a case study using the BE06 (Baker, 2009) and AmE06 (Potts and Baker, 2012) corpora, I first demonstrate how traditional concordance queries can be interpreted in a “fuzzy” way, allowing subtle differences in language usage to be captured without the need for careful crafting of the query itself. Next, I show how an embedding model can be used to cluster the results of a traditional concordance analysis based on semantic similarity, leading to novel groupings and orderings of results. I then show how an embedding model can be used to match expressions in one language variety with those in another, leading to truly novel concordance analyses. The paper finishes with a discussion of future directions in AI and the potential impact on concordance tool development.

References

Anthony, L. (2018). Visualization in corpus-based discourse studies. In C. Taylor & A. Marchi (Eds.), *Corpus approaches to discourse: A critical review* (pp. 197–224). Routledge.
Anthony, L. (2022). What can corpus software do? In A. O’Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 237–276). Routledge

Baker, P. (2009). The BE06 Corpus of British English and recent language change. *International Journal of Corpus Linguistics, 14*(3), 312–337.
Potts, A., & Baker. P. (2012). "Does semantic tagging identify cultural change in British and American English?" *International Journal of Corpus Linguistics 17*(3), 295–324.

Stephanie Evert (Friedrich-Alexander-Universität Erlangen-Nürnberg) ***A mathematical model of algorithms for organising concordances***

An important step towards achieving transparency in concordancing is aligning the hermeneutics of the process with the computational algorithms available to support it. To connect algorithms and their combinations to the interpretative part of concordance reading, we propose a formal framework that systematises classes of algorithms based on their mathematical properties and determines how different algorithms can be combined. Our framework categorises algorithms into five strategies based on how they manipulate the concordance view displayed to the analyst:

(1) *Selecting* algorithms subset concordance lines, typically using metadata categories or manual selection (e.g. ranges of lines, or one or more of the sets formed by a partitioning or clustering algorithm, see below).
(2) *Sorting* algorithms rearrange concordances by comparing pairs of lines (A, B) to determine whether A should sort before B, B before A, or both are tied. A typical example would be to sort alphabetically by the right or left context of the node.
(3) *Ranking* algorithms also rearrange concordances, based on a numerical value assigned to each line, with the largest values shown at the top of the concordance view. Examples include readability scores, average word frequency, or number of salient collocates in the context.
(4) *Partitioning* algorithms divide concordances into sets of lines that share a certain observable feature. Such sets could consist of all lines from the same text genre or author, all lines where the token immediately to the left of the node has the same POS tag, or lines that have been manually categorised according to bespoke criteria. The criteria by which lines are partitioned also provides frequency counts for the property of interest (= sizes of the sets).
(5) *Clustering* algorithms collect concordance lines into hierarchically nested sets based on their mutual similarity (with a flat list of clusters as a special case). Examples include flat clustering based on lexical overlap or semantic similarity and a "POS tree" display that groups lines by the POS tag of the token to the right of the node at the highest level, then by the tag of the second token to the right, etc. Mathematically, clustering is represented by an ordered tree whose nodes correspond to sets of concordance lines.

Multiple sorting and ranking algorithms can be combined: the second algorithm breaks ties in the ordering of the first, the third breaks any remaining ties, etc. By contrast, only a single partitioning or clustering algorithm can be in effect because of potential conflicts between sets formed by different algorithms. This single partitioning or clustering algorithm determines the high-level organisation of the concordance, while lines within each set are ordered according to the sorting and ranking algorithms. Selecting plays a special role: it allows the analyst to "zoom in" on part of a concordance for more fine-grained analysis and forms a natural scope boundary. In this way, multiple partitioning and clustering algorithms can be used together in an analysis path, one after each selecting step.

Acknowledgement: This work has partially been funded by the *Deutsche Forschungsgemeinschaft* (*DFG*, German Research Foundation) – 508235423.

Levi Dietsch and Alexander Piperski (Friedrich-Alexander-Universität Erlangen-Nürnberg)
FlexiConc demonstrator: A front-end web app for structured concordance analysis

FlexiConc is a software library developed to support a systematic approach to concordance analysis and interpretation by implementing existing and novel algorithms. It is not a comprehensive corpus management tool; rather, a 'concordance management tool' that can be integrated with other software. Testing and evaluating *FlexiConc* requires a front-end interface, which raises questions regarding the visualization of concordance reading strategies and the distribution of tasks between the front-end and back-end. In this talk, we will rationalize our design decisions and present a working version of the *FlexiConc* demonstrator.

The process begins when a user sends a query through the *FlexiConc* demonstrator to a host app, which could be any existing corpus management tool (e.g., *Corpus Workbench*, *CLiC*, *Sketch Engine*, *AntConc*). The host app returns the concordance data, which *FlexiConc* then passes to the library where users perform the required concordance operations.

Many corpus management tools (e.g., *CQPweb* and *Sketch Engine*) record procedural steps so that users can follow the sequence of operations performed and, if necessary, return to a previous step and continue from there. *FlexiConc* adopts a more intricate structure: the **operation-and-subset tree**, which facilitates complete research documentation. A set of concordance lines is represented as a node which can undergo various re-ordering (sorting and ranking), partitioning, and clustering operations. These are added as leaves attached to this node. Focusing on a subset of concordance lines—either through automatic selection or manual annotation—introduces a **scope boundary**. In terms of the tree, it is a node which can be further expanded with leaves by reapplying re-ordering, partitioning, and clustering operations. Nodes in the tree that are crucial for analysis can be marked as **snapshots** for later reproducible access by analysts or readers.

Figure 1 presents a prototype design for the *FlexiConc* demonstrator, including an operation-and-subset tree on the left. The current view (marked by an asterisk) selects lines from texts written in the 19th century and ranks them by the number of possessive pronouns in context.

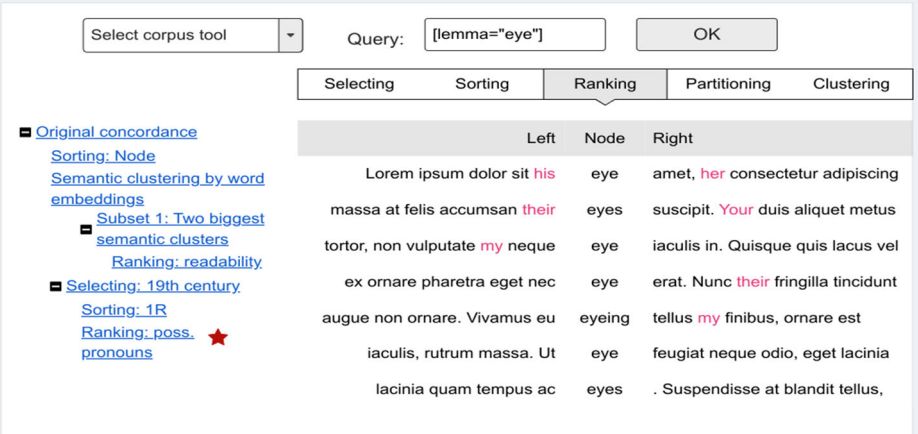

Figure 1. Prototype design for *FlexiConc* demonstrator.

The operation-and-subset tree effectively demonstrates how different concordance reading algorithms interact. When a user requests the application of an algorithm to a concordance view, two scenarios are possible:

- A child node is created from the current node (common when applying a Selecting algorithm).

- A sister node is formed, indicating either incompatibility with the current view or an override of the current algorithm. For example, Clustering algorithms are incompatible with each other; Ranking by readability, while technically compatible with Sorting by left context, adds new ordering scores with very few ties to the concordance lines, effectively overriding their previous order.

In summary, the purpose of the *FlexiConc* demonstrator is to illustrate a possible implementation of *FlexiConc* and present ways in which concordance analysis and interpretation can benefit from its features.

Acknowledgement: This work has partially been funded by the *Deutsche Forschungsgemeinschaft* (*DFG*, German Research Foundation) – 508235423.