RiCL Research in Corpus Linguistics

Review of Gillings, Mathew, Gerlinde Mautner and Paul Baker. 2023. *Corpus-Assisted Discourse Studies*. Cambridge: Cambridge University Press. ISBN: 978-1-009-16815-1. DOI: https://doi.org/10.1017/9781009168144

Tamsin Parnell
University of Nottingham / United Kingdom

Corpus linguistics and discourse analysis have long been said to exhibit a methodological synergy (Baker *et al.* 2008). The combination of approaches allows researchers to achieve both breadth and depth in their analysis while countering the criticism that discourse studies are prone to 'cherry-picking'. The burgeoning field of Corpus-Assisted Discourse Studies (CADS) is testament to the power of combining the two approaches to address contemporary social issues. However, as Gillings, Mautner, and Baker recognise, the uptake of the approach outside of linguistics "has not been as enthusiastic as might be expected" (p. 1). *Corpus-Assisted Discourse Studies* aims to redress this. The book offers a delicate balance between theoretical and empirical insights, peppered with relevant case studies that demonstrate how to conduct a corpus-assisted discourse study. Spanning seven chapters, the Cambridge Element provides beginners with a clearly explained introduction to the research area. As such, it would be appropriate not only for undergraduate and postgraduate students within linguistics, but anyone interested in the relationship between language and society.

Chapter 1 begins by explaining that CADS research explores discourse by examining corpora. It highlights areas of interest for the CADS researcher, including social representation, ideology, diachronicity, and institutional discourses, acknowledging that these are tied together by social questions rather than purely linguistic ones.

Chapter 2 —entitled 'The Rationale for CADS'— outlines the trajectory of CADS research, from the early linguistic interest in social questions (as pioneered by Firth) to Baker's (2006) seminal monograph *Using Corpora in Discourse Analysis*. According to the authors, part of the rationale for CADS is that it "puts analyses on more reliable empirical foundations" (p. 6). Going deeper, corpus linguistics and discourse analysis are united by a focus on linguistic patterning: combining the two approaches allows researchers to reveal the "incremental effect" of discourse (Baker 2006: 13). Noting that corpus linguistics allows both quantitative and qualitative insights, Gillings, Mautner, and Baker explain that the CADS researcher should oscillate between quantitative and qualitative components and can combine corpus linguistics and discourse studies in "any number of ways" (p. 8), as it is the combination of approaches that enables triangulation.

The third chapter leads the reader through the process of building a corpus for CADS research, starting by underscoring the importance of 'representativeness'. A distinction is made between reference corpora (typically representative of a broad language variety such as British English in the early 2010s) and specialised corpora (which represent a smaller language variety such as the works of Charles Dickens). Reference corpora, as the chapter explains, "provide an important benchmark against which [the discourse analyst] can interpret the evidence gleaned from their specialised, purpose-built corpora" (p. 9). In building the specialised corpus, what matters is that "the volume and the nature of the data are 'appropriate' for the research question" (p. 9). As the authors state, CADS researchers often work with newspaper data, with each article constituting a single text saved in txt format. Newspaper articles are popular texts with corpus linguists because they are not only politically significant but are easy to collect. Currently there are questions surrounding the collection of some other data types, including social media content. For example, is it ethical to combine a corpus of tweets when posters might not expect such public scrutiny? Here, the authors signpost to useful research, including Collins (2019) and Lutzky (2021). The final question the authors answer in this chapter is how big a corpus should be for CADS research. They explain that 'bigger' is not always better when it comes to CADS, and that the answer will depend on your research question.

Chapter 4 provides readers with a corpus toolkit, that is, a range of methods that can be used to answer your research question. The methods covered are frequency

analysis, concordance analysis, collocation analysis and keyword analysis. The explanation of frequency includes a helpful distinction between types and tokens, as well as an overview of tagging (both parts-of-speech and semantic). It elucidates the processes of creating a wordlist (ordering linguistic units either alphabetically or by frequency) and running searches for linguistic units across parts of a corpus (subcorpora). Case studies of UK Supreme Court judgements including at least one dissenting argument and a corpus of *Administrative Science Quarterly*[1] articles and book reviews usefully illustrate the power of frequency analysis to generate further questions. The importance of 'dispersion' is also touched upon, as words may be frequent in only one or two texts and therefore not be representative of the corpus as a whole, although this clustering may lead to further discourse analytical insights.

Section 4.2 covers concordance analysis, including important technical information such as how to sort, thin, and expand the concordance lines to make them easier to manage. It distinguishes CADS from other linguistic areas by explaining that, in this perspective, "discourse is the focus of analysis, and corpus assistance helps us to link large-scale social phenomena with linguistic choices at the micro level" (p. 23). To achieve this macro-level and micro-level synergy, researchers must go beyond the concordance line both in the sense of reading the co-text and considering the social context that shapes and is shaped by the corpus. The authors set out four ways to conduct a concordance analysis (pp. 23–25) along the axes of structured-unstructured and bottom-up-top-down, noting that when completing actual research, these types may overlap. They also encourage critical reflection on concordance analysis, a topic which is addressed in more detail in Gillings and Mautner (2024).

Section 4.3 addresses collocation analysis. Gillings, Mautner, and Baker reflect on how different methodological choices (such as length of collocational span) can alter results, encouraging experimentation to determine the most representative and useful set of collocates. A brief yet insightful discussion of statistical measures is offered in this subsection —an area which is a common cause of trepidation for those new to the more quantitative side to CADS. More sophisticated approaches to collocation analysis are also given a special mention, including the *Sketch Engine's Word Sketch* tool (Kilgarriff *et al.* 2014), and *#LancsBox*'s collocational network visualisations (Brezina *et al.* 2015). These are only cursory overviews, undoubtedly due to the audience and word limit.

[1] https://journals.sagepub.com/home/asq

Nevertheless, signposting to further reading that covers these areas would have been useful for those readers looking to progress to more complex approaches.

Section 4.4 introduces keyword analysis. Again, the discussion of techniques for calculating keyness is important for equipping new researchers with the confidence to choose which techniques to use. Equally enlightening is the explanation of how to group keywords and which to focus on. Perhaps the most important reflection, however, is that it is important to capture not just 'differences' between corpora, but also similarities. The authors establish ways to investigate similarity, including comparing two corpora against a third reference corpus.

Chapter 5 is titled 'CADS in Practice'. The strength of this chapter is its case study. Returning to the *UK Supreme Court* corpus explored in a previous case study, the authors present an analysis conducted on *Sketch Engine* in which they explore lexemes expected to play a part in expressing dissent. Their frequency study of *disagree* produces an "underwhelming result" (p. 40). To find a more fruitful result, they offer two approaches. The first one is via knowledge conventions about the genre of legal writing, which can be gained by reading a "fair number of texts from the corpus" (p. 40); this knowledge would lead us to the collocation *I disagree*. The second approach would be to look at the collocations of *disagree*, which reveal that the intervening adverb *respectfully* is more frequent in the dissenting subcorpus than in the majority subcorpus. The finding that "one of the characteristics of judges" framing of dissent is to buffer its impact with standardised politeness markers (p. 41) produces further research questions. Ultimately, the case study helpfully shows how different tools allow different routes into the data, and how promising paths can be distinguished from blind alleys (p. 42).

Another important facet of Chapter 5 is its recognition that CADS methods are seldom linear and orderly. Rather, "a little messiness" should be expected —albeit "without jettisoning the idea of systematic and transparent data analysis" (p. 43). To address the messiness, Gillings, Mautner, and Baker offer a musical metaphor in which each tool is regarded as an instrument. The point of the metaphor is to show that "CADS uses corpus tools flexibly, iteratively, and in a mutually reinforcing manner" (p. 44).

Chapter 6 discusses the limitations and potential pitfalls of CADS. The authors acknowledge that because CADS requires a "lexical hook," it is harder to identify

"broader discursive phenomena with multiple and unpredictable lexical realisations" (p. 45) such as argumentative strategies or extended metaphors. In this case, the researcher must return to discourse analysis proper. Equally, CADS can tell us little about "how meaning unfolds in longer stretches of text' and 'how interactants negotiate meaning in conversation" (p. 45). Thirdly, it is hard for CADS researchers to identify 'absences' in the data (although contrastive techniques can help to remedy this). Of course, there is also the issue of examining multimodal data through corpus linguistics methods, which —while increasingly taking place— is still difficult to do. Despite these issues, the authors question whether they can be referred to as limitations, since "CADS should be judged against what it was designed to do in the first place" (p. 46).

For beginners, an important aspect of the chapter is the section dedicated to pitfalls in CADS research. Drawing on their expertise as reviewers and seasoned CADS researchers, the authors remind readers that texts should not be collected just because they are easy (resulting in an 'all you can eat' approach), but because they are an integral part of the corpus. They also recommend discussing interpretations of data with colleagues to ensure it "passes the litmus test of intersubjective validity" (p. 48). Finally, the authors discuss the writing up stage and how necessary it is to strike a balance between reporting too much information about the methodological process and too little. This guidance is undoubtedly useful for those writing up a CADS project for the first time.

Chapter 7, which is the final chapter, reflects on the research journey. The authors acknowledge that their account of the research process has been selective and is necessarily incomplete (although, I would argue that it is sufficiently detailed to support beginners). In this chapter, the authors make the pertinent point that "disciplinary labels and identities ought to matter less than the commitment to unravel the mysteries of language" (p. 51), an important reminder for those who are working in interdisciplinary teams. They conclude by outlining areas in which CADS is developing, including how keywords are calculated, automating ways of categorising keywords, the use of *R*,[2] and research in languages other than English.

Overall, I would recommend *Corpus-Assisted Discourse Studies* to anyone interested in how a corpus linguistic approach to discourse analysis can strengthen research into social questions. I would encourage those tentatively reading this review

[2] https://www.r-project.org/

from outside of linguistics to take the leap and experiment with the tools outlined in the book. I would also suggest that those with more experience of CADS research read the book as a refresher, not least for the reflections on the pitfalls and potential limitations of the approach.

REFERENCES

Baker, Paul. 2006. *Using Corpora in Discourse Analysis*. London: Bloomsbury.

Baker, Paul, Costas Gabrielatos, Majid Khosravinik, Michał Krzyżanowski, Tony McEnery and Ruth Wodak. 2008. A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society* 19/3: 273–306.

Brezina, Vaclav, Tony McEnery and Stephen Wattam. 2015. Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics* 20/2: 139–173.

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

Gillings, Mathew and Gerlinde Mautner. 2024. Concordancing for CADS: Practical challenges and theoretical applications. *International Journal of Corpus Linguistics* 29/1: 34–58.

Kilgarriff, Adam, Vit Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý and Vít Suchomel. 2014. *The Sketch Engine*: Ten years on. *Lexicography* 1/1: 7–36.

Lutzky, Ursula. 2021. *The Discourse of Customer Service Tweets: Planes, Trains and Automated Text Analysis*. London: Bloomsbury.

Reviewed by
Tamsin Parnell
University of Nottingham
School of Cultures, Languages and Area Studies
Room B29a Trent Building
University Park
Nottingham
NG7 2RD
United Kingdom
E-mail: tamsin.parnell2@nottingham.ac.uk