Cambridge Elements

Corpus Linguistics

Corpus-Assisted Discourse Studies

Mathew Gillings
Gerlinde Mautner
Paul Baker

ISSN 2632-8097 (online)
ISSN 2632-8089 (print)

Cambridge Elements

Elements in Corpus Linguistics
edited by
Susan Hunston
University of Birmingham

CORPUS-ASSISTED DISCOURSE STUDIES

Mathew Gillings
Vienna University of Economics and Business
Gerlinde Mautner
Vienna University of Economics and Business
Paul Baker
Lancaster University

CAMBRIDGE
UNIVERSITY PRESS

https://doi.org/10.1017/9781009168144 Published online by Cambridge University Press

CAMBRIDGE
UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8EA, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre,
New Delhi – 110025, India

103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment,
a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of
education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781009168151

DOI: 10.1017/9781009168144

First published 2023

A catalogue record for this publication is available from the British Library.

ISBN 978-1-009-16815-1 Paperback
ISSN 2632-8097 (online)
ISSN 2632-8089 (print)

Corpus-Assisted Discourse Studies

Elements in Corpus Linguistics

DOI: 10.1017/9781009168144
First published online: March 2023

Mathew Gillings
Vienna University of Economics and Business

Gerlinde Mautner
Vienna University of Economics and Business

Paul Baker
Lancaster University

Author for correspondence: Mathew Gillings, mathew.gillings@wu.ac.at

Abstract: The breadth and spread of corpus-assisted discourse studies (CADS) indicate its usefulness for exploring language use within a social context. However, its theoretical foundations, limitations, and epistemological implications must be considered so that we can adjust our research designs accordingly. This Element offers a compact guide to which corpus linguistic tools are available and how they can contribute to finding out more about discourse. It will appeal to researchers both new and experienced, within the CADS community and beyond.

Keywords: CADS, corpus-assisted discourse studies, discourse analysis, corpus linguistics, methodology

ISBNs: 9781009168151 (PB), 9781009168144 (OC)
ISSNs: 2632-8097 (online), 2632-8089 (print)

Contents

https://doi.org/10.1017/9781009168144 Published online by Cambridge University Press

1 Introduction

Corpus-assisted discourse studies (CADS) explores discourse (i.e., language as social practice) through examining corpora (i.e., large computerised sets of textual data). CADS allows one to survey a corpus in its entirety rather than focusing only on certain texts which, by accident or design, may be those that confirm what one wanted to show all along. It thus counteracts the 'cherry-picking' charge that has often been levelled at discourse studies (DS). Researchers also rightly appreciate the ease with which corpus linguistics (CL) can be integrated into different types of research designs; the new perspectives thereby opened up; the boost to both the depth and breadth of the analysis; and the greater confidence it gives the analyst in interpreting the results.

1.1 Who This Element Is For

This Element is for you if your research uses large amounts of language data that you wish to approach from a discourse perspective. Although *discourse* is a notoriously fuzzy term (Stubbs, 1983; Baker and Ellece, 2011; Mautner, 2016: 16–24), used in different disciplines and with various shades of meaning, it has a fairly robust semantic core consisting of three key components. *Discourse* refers to (i) longer stretches of language (usually complete texts and interactions rather than merely single sentences) which (ii) occur naturally in a specific social context and (iii) are analysed as performing social functions. Typical questions that a discourse analyst might ask are: How is language used to represent a particular social group? Which linguistic choices correspond with which ideological position? Do discursive representations change over time? What role do particular linguistic choices play in institutional discourses? In answering these types of questions and many more, CADS is a useful ally.

What all CADS projects have in common is that they have a social question at their centre rather than a purely linguistic one. That question may involve an issue such as inequality, poverty, racism, or other social ills. But projects may equally be driven by a more general interest in the links between a social practice and its associated linguistic choices. CADS can make a useful contribution to unpacking what makes discourse tick, which is why it should be an attractive option not only for linguists but also for any researcher investigating the interplay between discourse and society, whether they work in sociology, psychology, law, management, or indeed any other discipline with an interest in discourse. That said, the actual uptake of the approach outside linguistics has not been as enthusiastic as might be expected, and we thus hope that by presenting this Element, we can further encourage discourse scholars to the field who may not have given it much thought so far. It is also for this reason that we write mainly about the corpus

https://doi.org/10.1017/9781009168144 Published online by Cambridge University Press

linguistic side of the analysis, and assume prior knowledge of the more widely used discourse analytical methods.

Discourse data could come from any number of sources, genres, or media: newspaper articles, political speeches, social media postings, or recordings of focus group interactions, to name just a few. For our present purposes, it is also immaterial which theoretical or methodological tradition you were socialised into and which research paradigm you prefer. As long as your data consists of language, and you accept that language and society influence each other, CADS will be a promising route to follow, and this Element should help you add a few items to your toolbox. The Element has been written primarily for research students, particularly those encountering CADS at Master's or PhD level. However, if you are already an experienced CADS researcher and keen to introduce the method to your students, you may find the Element useful as a compact teaching resource.

1.2 Aim and Structure of the Element

Although there is no shortage of published work using CADS, it would be challenging for anyone new to the field to find all the essential information in one place. This is the gap the present Element aims to fill, with the following two goals in mind: (i) to provide a succinct 'how-to' guide for researchers wishing to use CADS, allowing them to carry out their own analyses with as much rigour as possible; and (ii) to identify the limitations of this approach and encourage researchers to critically reflect not only on the method but also on their own biases and interpretations.

In Section 2, we discuss the rationale for combining CL and discourse analysis. Section 3 talks about corpus building for CADS projects. Section 4 describes the four main corpus tools, each in their own subsection: frequency, concordance analysis, collocation analysis, and keyword analysis. We aim to get back to basics here, starting from the beginning and exploring how these processes can give us an insight into discourse. We will use a range of examples from our own work to illustrate how this is achieved, making it as easy as possible for readers to get started with 'doing' corpus-assisted discourse analysis. Each of the four subsections will address the following questions: What does the tool do? How can it be used for studying discourse? What are things to watch out for when applying it? Throughout, we will emphasise the importance of being critical of one's own practice, rather than blindly accepting numbers and other outputs at face value.

Section 5 aims for a more integrated picture, starting with a worked example to show how the different tools can be made to interact, and then introducing

https://doi.org/10.1017/9781009168144 Published online by Cambridge University Press

a musical metaphor, a new model for theorising this interplay between different tools. Section 6 discusses some of the main limitations of corpus-assisted discourse analysis. Building on discussions by Taylor and Marchi (2018) and Mautner (2015, 2022), the section focuses on the method's potential shortcomings. These will be discussed against the background of more general questions around epistemology, research design, and triangulation. Finally, in Section 7, we review the research journey that scholars embracing CADS typically embark on. We briefly address the challenges involved in mixed-methods research, reflect on the careful balance between intuitive and planned approaches to the data, and discuss how a 'craft attitude' to research helps one to make sense of messy data.

We expect our Element to be complementary to *Doing Linguistics with a Corpus* (Egbert, Larsson and Biber, 2020). Whereas that publication focuses strongly on quantitative aspects of the field, we will mainly explore the more qualitative and social angle. As Cambridge Elements are intended to be brief, focused, and accessible, we were forced to compromise on the amount of material and level of detail covered. Inevitably, some readers will identify gaps which they would have preferred to see filled. Whenever space constraints meant we could only scratch the surface, we hope that the references provided will direct readers to useful sources for further study. We also wish to point out here that we focus largely on English language CADS and the tradition in British linguistics. We will be presenting English data and mainly citing works by scholars from English-speaking countries. This is as a result of our own socialisation into linguistics generally and CADS specifically, but should by no means be taken as the only possible and worthwhile perspective.

2 The Rationale for CADS

The field of corpus linguistics (CL) has made great strides over the past five or six decades. It began as a methodological framework mainly applied to lexicography and language learning, but researchers soon began to see its potential for analysing discourse in social context. By the mid-1990s, the time was ripe for CL to team up with discourse analysis, and particularly with its critical variety, critical discourse analysis (CDA). Corpus-assisted discourse studies (CADS) was born and quickly became a field in its own right. Now it boasts the usual paraphernalia of consolidation (Mautner, 2019, 2022), such as regular conferences, monographs, handbooks (e.g., Friginal and Hardy, 2021), and a dedicated journal, the *Journal of Corpora and Discourse Studies*[1] (though papers based on CADS of course also continue to be published in other

[1] https://jcads.cardiffuniversitypress.org.

https://doi.org/10.1017/9781009168144 Published online by Cambridge University Press

journals). In this section, we will take a look at CADS' historical roots and present concerns.

2.1 CADS: Past and Present

As a field, linguistics became interested in social questions long before CADS, as such, entered the scene. Firth's pioneering work is clearly pivotal here; he expressed an interest in 'the detailed contextual distribution of sociologically important words' (Firth, 1935: 40), an early harbinger of corpus and discourse studies (DS) joining forces. Sixty years later, CADS was first put on the map by publications such as Leech and Fallon (1992), Stubbs and Gerbig (1993), Caldas-Coulthard (1993), Hardt-Mautner (1995), and Stubbs (1997). It is also now more than fifteen years since the release of Baker's 2006 book, *Using Corpora in Discourse Analysis*, though it is still considered to be a go-to guide for new and experienced researchers alike. The present Element builds on his work, while also taking into account advances in both the available technology and the associated techniques.

Since the 1990s, when CADS took off, important meta-level methodological and epistemological questions have been addressed, surrounding reflexivity, bias, and triangulation (Marchi and Taylor, 2009; Mautner, 2015; Baker and Egbert, 2016; Taylor and Marchi, 2018; Egbert and Baker, 2020). These continue to be relevant not only in theoretical terms but also because of their practical implications for anyone developing a research design and wondering which method(s) to choose. That choice is as crucial as it is difficult, particularly if a project crosses disciplinary boundaries. As a newcomer to CADS or linguistics generally, you may be attracted to exploring your data through a language lens, but may not be aware of the methods that are on offer. Nor may you have the time or inclination to track nuanced meta-level methodological discussions over a series of individual papers and book chapters. With this need in mind, our Element also aims to create a one-stop shop for people encountering CADS for the first time.

Predictably, the maturity of the field has also entailed a degree of diversification. There are now different variants of CADS around, depending on whether they are located more at the CL end or at the DS end of the spectrum. A range of labels has sprung up, including *corpus-led*, *corpus-oriented*, *corpus-informed*, *corpus-based*, and *corpus-driven*, with the latter pair having attracted notable and widely received theorising (Tognini-Bonelli, 2001; McEnery and Hardie, 2012; Biber, 2015). Yet, however strongly individual authors – or indeed some of our readers – may feel about any substantive differences that these terms imply, we would argue that for our purposes here there is no need to explore this

https://doi.org/10.1017/9781009168144 Published online by Cambridge University Press

variety any further. After all, no matter how it is sliced or diced, the same corpus tools are on offer. In this Element, we treat *corpus-assisted* as an umbrella term. We primarily see CADS as the examination of textual data, applying a corpus linguistic methodology, to explore the two-way relationship between discourse and society. This can come in different shades, of course, but in the present context we tend to focus on questions around discursive representation rather than, for example, contrastive analyses of text types.

The variety of projects to which CADS has been usefully applied is impressive, ranging from business (e.g., Koller, 2007; Lischinsky, 2011; Fuoli, 2018; Jaworska, 2018; Lutzky, 2021a) and law (e.g., Tkacukova, 2015; Phillips and Egbert, 2017; Solan and Gales, 2017; Wright, 2017; Gillings, 2022) to healthcare (e.g., Semino et al., 2018; Baker, Brookes and Evans, 2019; Hunt and Brookes, 2020), gender (e.g., Baker, 2014; Hunt, 2015; Formato, 2019; Zottola, 2021), and history (Gupta, 2015; McEnery and Baker, 2022; Taylor, 2022); see Nartey and Mwinlaaru (2019) for a systematic meta-analysis. There is corresponding variety in the types of corpora that CADS can be used with. Some studies work with reference corpora which are designed to represent a whole language, whereas others are based on specialised corpora that are built ad hoc, because the discourse type concerned is so specific that one cannot rely on readily available material (Partington, Duguid and Taylor, 2013; also see Section 3).

2.2 Flexible Synergies

Over the past few decades, qualitative discourse analysis has produced a wealth of insightful work advancing our understanding of linguistic patterns above the sentence level (Jaworski and Coupland, 2014), and shedding light on the mutual relationship between language and society. When coupled with an explicitly critical stance, research on discourse also becomes relevant in political terms. Appearing first as *critical linguistics*, and then developing into *CDA* and *Critical Discourse Studies*, the approach focuses on how discourse enacts ideology and power, and specifically, how discourse is implicated in creating and sustaining unequal power relationships, disadvantage, and discrimination. Landmark publications include Fairclough (1992), van Dijk (1993), and Wodak and Meyer (2015). Researchers in this tradition frequently adopt an ‘unabashedly normative’ stance (van Dijk, 1993: 253), committed not only to investigating social ills but also to changing them. The main interest of CDA remains language, but it typically mines disciplines other than linguistics (such as sociology and political science) for theoretical inspiration. Critical discourse analysis uses a purposefully eclectic set of qualitative tools and procedures. Its descriptive apparatus targets phenomena at all linguistic levels, ranging from large-scale argumentative

https://doi.org/10.1017/9781009168144 Published online by Cambridge University Press

strategies to details of word choice and grammatical patterning. Its analyses are based on close reading, thick (i.e., very detailed) description, and hermeneutic interpretation.

And therein lies the problem. Realistically speaking, close reading and thick description are possible only when the corpus is fairly small: a few newspaper articles, say, or a handful of transcripts. It is true that more material can be covered when software for qualitative content analysis is employed (such as AtlasTI, MaxQDA, or NVivo), but the necessary demands on person-power, time, and money can still be prohibitive. So the obvious solution is to use smaller, more manageable datasets. There is no question that these have yielded many worthwhile results in the past, and will continue to do so in the future. Yet, the smaller the corpus, the larger looms the question of representativeness and selection bias. How certain are we that our chosen examples are typical, so that we can confidently generalise our findings beyond the actual texts investigated? And what were our criteria for choosing these specific texts in the first place? Did they perhaps, as critics of CDA often claim, attract our attention precisely because they seemed to provide the very evidence that the analysis was meant to uncover? Is the argument therefore irredeemably circular? This is where CL comes in. Because it allows us to investigate a much larger number of systematically sampled texts, it puts analyses on more reliable empirical foundations. Size does not necessarily matter, but it inspires confidence.

Yet the rationale for CADS runs even deeper. We often think of CADS as a way of uniting *different* perspectives (i.e., the corpus and the discourse view). But what if we also thought of CADS as uniting *similar* perspectives? Here, the key link is linguistic patterning. It is both the home territory of CL and of central concern in DS. Corpus linguistics identifies regularities in the evaluative load of word partnerships; DS explains how these are related systematically to the sociopolitical context. Together, they can provide a compelling account of how discourses solidify through repeated, incremental usage (Stubbs, 2001: 215; Baker, 2006: 13).

Additionally, a great deal of mileage is to be had from CL's heuristic potential – that is, its ability to help us discover unexpected things. Certain computations may not only be interesting in their own right, but may also provide valuable clues about promising lines of inquiry, possibly even leading to research questions that originally weren't even on the agenda (Subtirelu and Baker, 2018: 108). To this end, researchers typically begin by simply playing around with the data, for example computing frequency lists and *n*-grams (see Section 4.1), and comparing them with those derived from other corpora. Even at this early, loosely structured stage, CL often gives one the kind of handle on the data that would elude manual analysis. Nonetheless, CL should not rashly be

https://doi.org/10.1017/9781009168144 Published online by Cambridge University Press

characterised as a purely quantitative approach (and hailed or dismissed as a result, depending on one's methodological preferences). It certainly has many strong quantitative components, but its main strength is that it enables the researcher to view large corpora through *both* a quantitative lens and a qualitative lens. Both views are needed if CADS is to be rigorous and relevant. And ideally, the time-honoured 'quant-qual divide' will be bridged as a result. That divide is substantive, of course, yet only partly so; to a significant extent it seems a construct of convention, an entrenched dichotomy separating not only methods but also mindsets.

CADS, by contrast, requires as much commitment to the computer-assisted profiling of corpora as to the human-led investigation of those discursive phenomena which (at present at least) are beyond the reach of automated analysis. Yet many works on CADS, including this Element, focus on the benefits of computer assistance while taking for granted the equally important, though more traditional, discourse perspective. Although the corpus-assisted angle seems comparatively novel and exciting, it is important not to become so enchanted by it as to forget that CADS is ultimately about finding out how *discourse* works. In this context, Ancarno (2020: 174–5) argues against a polarised, dualistic view of the quantitative–qualitative debate. She makes the very pertinent point that one should look beyond the standard narrative of CADS overcoming the deficiencies of qualitative discourse analysis. After all, the corpus linguistic perspective can and should also be enriched by the discourse element of CADS. So it works both ways. And indeed the most insightful projects combine CL's ability to uncover grammatical and semantic patterns in large corpora with the potential of DS for unravelling complex meaning-making processes in coherent stretches of text. To create added synergistic value, the analysis ought to involve 'oscillating' (Mautner, 2007: 66) or 'shunting' (Partington and Marchi, 2015: 231) between its quantitative and qualitative components.

The basic idea behind CADS, then, is to achieve a 'useful synergy' between CL and DS (Baker et al., 2008; Subtirelu and Baker, 2018: 107). There are various ways in which research designs can deliver on this promise, depending on when CL and DS, respectively, are brought on board and how they are made to interact. There are two prototypical approaches. Corpus linguistics can go first, leading to results which DS then helps interpret on the basis of detailed textual analyses as well as against the backdrop of historical, sociocultural, and political knowledge. Alternatively, DS can take the lead, identifying instances of discursive struggle in a small number of texts and then using CL to find out how typical these instances are across larger corpora. These two approaches are only the end points of a continuum; in reality, research designs can combine CL

https://doi.org/10.1017/9781009168144 Published online by Cambridge University Press

and DS in any number of ways. It is the combination, as such, that makes CADS inherently triangulated.

Yet what is it about triangulation that makes it so valuable? Before addressing this question, let us briefly do some conceptual groundwork. According to Marchi and Taylor (2009: 4–6), who draw upon Denzin (1970), there are four types of triangulation: investigator triangulation, using more than one researcher to explore the data; data triangulation, collecting data through several sampling strategies; theoretical triangulation, exploring data through more than one theoretical lens; and methodological triangulation (which itself can be split into *between-method* and *mixed-method* triangulations), using more than one method to collect and analyse data. Because CADS is a 'mixture' of CL and DS, and the two are interdependent, it falls into the mixed-method category, but it can be combined with any or all of the other types. Essentially, the value of triangulation lies in opening up different perspectives which would be closed to one researcher, dataset, theory, or method alone. A triangulated design thus helps validate and enrich the analysis.

In research designs based on triangulation, CADS is a valuable yet unassuming partner, making a distinctive contribution without irrevocably committing the researcher to a particular theory-cum-method package. On the contrary, CADS sits comfortably with different theoretical frameworks (as long as they are premised on the mutual relationship between language and society). It can be used together with the methods associated with those frameworks, increasing their leverage and making them more efficient. Or it can do something those other methods cannot do at all. What is more, CADS is malleable and does not tie researchers down to a particular identity. In order to 'do' CADS, you do not have to 'be' CADS. It is a set of methods, not a religious order.

3 Corpus Building for CADS

If Section 2 was successful in whetting your appetite for corpus-assisted discourse studies (CADS), then you may be wondering how to get started with such a project yourself. As the name suggests, to get started with a CADS investigation, you first need a corpus. Just like any dataset, corpora come in all shapes and sizes. Yet, according to formal definitions (which not all researchers choose to follow), a 'dataset' and a 'corpus' are not in fact the same. Both do indeed refer to collections of (language) data that can be subject to quantitative and qualitative analyses, but a corpus differs in that it relies firmly on the notion of *representativeness*. Rather than pulling together language data from anywhere and everywhere, corpora are carefully crafted with the overall aim of representing a particular language variety. Still, there are scholars who

https://doi.org/10.1017/9781009168144 Published online by Cambridge University Press

choose to use the term *corpus* more loosely – simply using it to refer to the collection of texts under analysis.

A corpus could be something as broad as the British National Corpus 2014 (BNC2014) (Love et al., 2017; Brezina, Hawtin and McEnery, 2021), which aims to represent British English as it stood throughout the early 2010s; but it could equally have more modest ambitions and aim to represent a smaller and more specialised variety – something like the works of Shakespeare (Culpeper et al., 2021) or the language of business meetings (Handford, 2010). These smaller corpora are incredibly useful to the discourse analyst, allowing 'a much closer link between the corpus and the contexts in which the texts in the corpus were produced' (Koester, 2022: 49). In this section, we will offer a brief introduction to two types of corpora – reference corpora and specialised corpora – and where we can find them.

Reference corpora are also known as *general* or *balanced* corpora. It is not every day that new ones are built; the process is time-consuming and expensive, and requires a great deal of planning, coordination, and person-power. Those that are available are frequently accessed by researchers to test hypotheses about language in general. They are commonly used by those interested in language teaching to help identify common lexicogrammatical patterns, and by corpus linguists as a point of comparison in keyword analyses (see Section 4.4). For the discourse analyst, reference corpora provide an important benchmark against which they can interpret the evidence gleaned from their specialised purpose-built corpora. Table A1 in the Appendix lists some of the places where reference corpora can be found.

Two of the most popular reference corpora for British and American English are, respectively, the aforementioned BNC2014 and the Corpus of Contemporary American English (COCA) (Davies, 2008). And because reference corpora are designed to represent a whole language, researchers often make claims about 'present-day English', as a whole, based on them. But we would be remiss, especially in a publication such as this, not to comment on their limitations. For one, a reference corpus is only as good as the individual texts that make it up. The BNC2014 includes spoken language, as well as various written texts (newspapers, fiction, and magazines) and even e-language; but deciding on text sampling proportions in the first place is inherently a subjective decision taken by the respective team responsible for its construction. Hence, its ontological status should not be oversold.

A second issue concerns the extent to which a reference corpus reflects the language users that it claims to represent. How the corpus is sampled can ultimately impact on how representative it is. For example, when developing a corpus of spoken English, one might consult census data and go for a stratified

https://doi.org/10.1017/9781009168144 Published online by Cambridge University Press

sample of texts, ensuring that there is a proportionate amount of text for each social stratum. Alternatively, one might prefer to randomly sample texts, or opt for mere convenience sampling – that is, taking what you can get. In addition, some of the questions involved may be sociological or even 'philosophical' in nature: How can we define what 'British English' is? Is it a question of speaker, genre, or context? In a globalised world, does defining language varieties in this way even matter? Categorising language varieties is not easy, which makes it all the more important for the sampling frame (i.e., the units to be sampled) to be clearly defined so that a corpus has the best chance at being representative (see Biber, 1993; Hunston, 2002; Love, 2020; Reppen, 2022).

For some scholars, and some research questions, a reference corpus may be all that is needed, especially when asking questions about the links between language and society at large. But for other questions, where the focus is on a specific language variety or genre, it may be necessary to build your own corpus from scratch. This is referred to as a *specialised* corpus. As Partington, Duguid and Taylor (2013: 12) rightly comment: 'CADS is [. . .] typically characterized by the "ad hoc" compilation of specialised corpora, since very frequently there exists no previously available collection of the discourse type in question.' Drawing upon the notion of representativeness, one should start by considering exactly which corpus would be useful to best answer the research question at hand. If one is interested in, say, the development of climate change discourse in the UK press (as in Gillings and Dayrell, 2023), then it might be enough to access the newspaper component of the BNC2014. But it might not – perhaps because there is not enough data, or perhaps because the data that is available is not specific enough to the topic concerned. That is, both the volume and the nature of the data have to be appropriate to the research question under investigation. It might make more sense, then, to collect a brand new corpus which consists only of those articles which have something to do with climate change. In doing so, you would have the flexibility to decide exactly which newspapers to include (broadsheets, tabloids, or both?), along with which time period is most useful to explore: a single period (for synchronic analysis), or several periods (for diachronic comparison). This is not only a process of data collection but also a process of 'reducing the breadth of your inquiry, while at the same time sharpening its focus' (Mautner, 2019: 8). The danger throughout this process is that there is a margin of error, and one should be careful not to select texts for a corpus in such a biased way that it gives only the expected results. If you did that, then you would rightly be accused of exactly the type of cherry-picking that a computer-assisted approach is meant to avoid.

Regarding newspaper corpora, Gabrielatos (2007) offers practical guidance on how texts can be systematically collected by pre-determining query terms,

https://doi.org/10.1017/9781009168144 Published online by Cambridge University Press

searching for articles in a text archive (such as LexisNexis or Factiva) and then downloading them. Typically, and following recommendations by Egbert and Schnur (2018) and Biber (2021) to treat the text as the main unit of analysis, this means that we construct the corpus article by article, saving one newspaper article per txt file. But even after collecting the raw texts, preparing the data for corpus analysis is another matter altogether. Often when converting text from one format to another, problems can arise that need fixing. These problems are referred to as *noise*, and may consist of superfluous spaces added between letters, punctuation marks that have been changed or lost, and so on (McEnery and Hardie, 2012; Joulain-Jay, 2017). And, depending on the intended uses, these files may need some form of XML mark-up or annotation to identify structural elements like section or paragraph breaks (McEnery and Hardie, 2012: 30). Of course, even if you decide against adding mark-up, you can still work with raw text in basic txt files which can then be uploaded to your choice of corpus analysis software, ready for analysis.

Because of their political significance, newspapers are a fairly common source of data within CADS (Cheng and Lam, 2013; Thornborrow, Ekstrom and Patrona, 2021; Räikkönen, 2022), with an added bonus being the relative ease with which articles can be collected. Other text types analysed in CADS include interviews (Dayrell, Ram-Prasad and Griffith-Dickson, 2020), parliamentary debates (Baker and Love, 2015; Appleton, 2021), oral histories (Fitzgerald, 2020), tweets (Harvey, 2020; Lutzky, 2021a, b), and Reddit posts (Dayter and Messerli, 2022; Krendel, McGlashan and Koller, 2022). In the latter three examples in particular, where social media data is concerned, not only must the researcher pay attention to the methodological issues outlined above, but to further ethical and legal problems. Questions currently being asked include: Is it ethically sound to compile a corpus of tweets? How public do tweet authors expect their messages to become? And more generally, how is 'the public domain' defined? Should corpus linguistics (CL) also require informed consent? Is anonymisation enough? Where does the law stand? What part can university ethics panels play? These questions, and more besides, are discussed in Lutzky (2021a), along with Collins (2019) and BAAL (2021).

And finally, what about spoken language? If one is interested in a particular form of spoken discourse, then compiling a purpose-built spoken corpus may be necessary. This is challenging. Not impossible, of course, but generally, time and funding is needed to achieve the intended result. The compilation of spoken corpora requires someone to get access to participants, gain informed consent, and then make the audio recordings using first-rate equipment (although see Knight et al. (2021) for an overview of how crowdsourcing methods were used to create the National Corpus of Contemporary Welsh). After collecting the data, the researcher (or, more likely, a team of researchers) must then listen to those recordings and transcribe them according to pre-determined and agreed-upon guidelines. And these guidelines come with differing levels of complexity. Some researchers will definitely want to include, for example, pauses, interruptions, and overlaps, whereas for others that level of detail would not only not be necessary, but in fact also distracting from their core research question. While the compilation of spoken corpora is generally more resource-intensive, the result may be that the researcher is much closer to the data, having already spent so long on corpus construction, and is thus able to interpret discourses more effectively.

Before we close this section, we should address one final commonly asked question: How large should a corpus be? Or, in other words: When can we be confident that we can stop collecting data? In true academic style, we would hedge and say that ‘it depends’. Bigger is not always better when it comes to CADS, and broadly speaking, the more homogenous the corpus, the smaller it can be. The answer rather lies in how much data is necessary to answer your research questions. To use a simple example from Mautner (2019: 9):

> If one wants to make claims about annual reports of a company that has only existed for five years, then gathering those five annual reports makes a corpus 100% representative. If, however, the company has been around for 50 years, then five annual reports are less useful. It is important to note that the concept of saturation typically invoked in sampling – where one stops collecting data when it becomes more of the same – does not quite work within CADS. In fact, ‘more of the same’ is not mere redundancy, but it shows us exactly the cumulative effect of repeated linguistic choices.

We do not have the space within this Element for a comprehensive step-by-step guide on corpus building, but we can recommend McEnery and Xiao (2006), Reppen (2022), Knight and Adolphs (2022), and McEnery and Brookes (2022) as good sources. Instead, we have outlined the key principles of corpus building – a brief overview of what should enter the thought process. Understandably, much of this may be too much for a single researcher to handle alone. If you are in that position, help is at hand. Firstly, many corpus analysis programs, such as CQPweb (Hardie, 2012), Sketch Engine (Kilgarriff et al., 2014), and #LancsBox (Brezina, McEnery and Wattam, 2015), come equipped with pre-made corpora. Some of these corpora are the aforementioned reference corpora, but others are more specialised. And secondly, colleagues working in other academic fields might be able to provide data. For interdisciplinary collaboration to come into its own, however, the dialogue between colleagues from different fields needs to get underway as early as possible in the course of the project so that agreement can be reached about how to process and format

https://doi.org/10.1017/9781009168144 Published online by Cambridge University Press

data. (Collins and Hardie (2022) make a series of useful suggestions for how mark-up can be standardised and made unambiguous throughout the corpus.) As we noted in Section 1, linguistics is far from the only field to study discourse data, and colleagues across the academy may be unaware of the treasure chest that they are sitting on.

Key takeaways and things to watch out for:

- Decisions made at the corpus design, sampling, and text preparation stage have a direct implication on how you will be able to conduct your analysis, and the results you will be able to derive.
- Reference corpora are generally representative of whole language varieties and are useful for exploring how language operates in real-life contexts.
- Specialised purpose-built corpora tend to be the most widely used in CADS, but care must be taken not to sample texts in such a biased way that the analysis offers only the results you expect to find.
- When deciding what size a corpus ought to be, bigger is not always better. Instead, what matters is how suitable the corpus is for answering certain research questions.
- Pre-made corpora, available via corpus analysis programs, are very useful resources and efficient to use.
- Exploring discoursal data not originally intended for CADS, perhaps through interdisciplinary partnerships, may yield fruitful outcomes for all parties involved. To ensure effective collaboration, however, an understanding has to be reached early on about how the data is to be formatted and marked-up.

4 A Corpus Linguistic Toolkit for Studying Discourse

After laying out a set of research questions, and accessing, acquiring, or building a corpus, the next phase of the research process can begin. As you sit down at your laptop, ready to start the analysis, you may be thinking: What now? How do I actually *do* corpus-assisted discourse analysis?

This is the question that the present section aims to answer. Here, we present four subsections that are each dedicated to one of the four main corpus linguistic tools: frequency, concordance analysis, collocation analysis, and keyword analysis. Each subsection aims to answer the following questions: What does the specific tool do? How can it be used for studying discourse? And what are

https://doi.org/10.1017/9781009168144 Published online by Cambridge University Press

things to watch out for when applying it? To answer these questions, we will draw upon several illustrations and examples from our own work.

What will become immediately clear is that not all of these tools are used for all analyses. While it is true that most work within corpus-assisted discourse studies (CADS) will make use of frequency information, and it is also true that discourse analyses tend to return to the data through concordance lines, not all of them will feel the need to explore collocations or perform a keyword analysis. Sometimes it may be appropriate to compare your corpus with another, to find out what is most typical of that corpus, but sometimes you may wish to just look internally. The extent to which these techniques are employed differs depending on the research questions. In Section 5, after our introduction to the four main techniques, we will illustrate how they work together.

There are many corpus analysis programs that can be used to aid these analyses. Some of these programs are commercially produced, whereas others are freely accessible; some of them are online-only, whereas others require a download. Different researchers tend to have different allegiances to these programs, largely dependent on which they were taught first, but also on whether funding is available, the (perceived) ease-of-use, and the availability of specific tools within the program concerned. McEnery and Hardie (2012) offer a brief history of these programs, detailing how they started as simple ‘Key Word in Context’ displays, before expanding to incorporate more and more tools. In the Appendix, we list some of the most well-known and most frequently used programs used for analyses within CADS. This list is not comprehensive, but it instead offers a taste of what is out there.

4.1 Frequency

Frequency information is not only at the heart of corpus linguistics (CL) generally, but it is also important specifically within CADS projects. In due course, we will introduce and explain a few key concepts essential for working with frequency data; these include *wordlist*, *tokenisation*, and *n-grams* (see also Evison (2010) for an overview).

All of the most commonly used corpus analysis programs (see Appendix) have the ability to calculate frequencies, and thus allow the user to create a wordlist (i.e., a list of all the words contained in the corpus, ranked alphabetically or by frequency). Exactly how that frequency is calculated, however, is directly linked to how the concept of a ‘word’ is defined. Sometimes, it might be enough to count words in the same way as a simple word processor: that is, by counting the number of linguistic items that have white space on either side. But counting words in this way comes with a whole gamut of issues: How do we treat

https://doi.org/10.1017/9781009168144 Published online by Cambridge University Press

contractions (e.g., *don't*) or hyphenated words (e.g., *great-grandmother*)? How do we treat punctuation marks? Should they be searchable in their own right?

These questions suggest that it might be better to count words in another way. We could choose to break a text down into individual tokens (a single meaningful linguistic unit that is processed by the corpus analysis program). For example, we might find it useful to treat the word *she's* as two separate tokens (*she* + *'s*), as the third person pronoun *she* is a meaningful linguistic unit, in addition to *'s* as a contraction of *is*. The process of splitting words up in a pre-defined manner is called *tokenisation*. By tokenising words, the user has the flexibility to decide exactly which units should count as individual tokens, and can therefore run searches for these linguistic units on their own.

For most off-the-shelf corpus analysis programs, the method of tokenisation is stipulated by the developers and generally occurs when the user imports their corpus.[2] At that stage, some tools also tag the corpus. *Tagging* refers to the process of adding additional information, such as assigning a grammatical or semantic category to each word. It is now fairly commonplace for corpora to undergo a form of automatic – and indeed very reliable – grammatical part-of-speech tagging, so that the researcher can distinguish between, say, the modal verb *might* and the noun *might*. CLAWS, one of the most commonly used taggers, has an accuracy of 96–97 per cent, with 'the precise degree of accuracy varying according to the type of text'.[3]

Semantic tagging, on the other hand, allows the researcher to identify connections based broadly on meaning rather than grammar or direct lexical relationships. The idea is to be able to explore the spread of different semantic domains across the corpus. One of the most well-known semantic taggers is USAS, which divides language into 21 broad discourse fields (i.e., essentially, semantic domains), which can be expanded into 232 more fine-grained category labels.[4] Examples of such fields are *Money and commerce*, *Emotional actions*, *states and processes*, and *Food and farming*. If the corpus under analysis is tagged for semantic category, the researcher can explore thematic differences across datasets, rather than being restricted to lexical or grammatical searches alone (see Potts and Baker (2012) and Potts (2015) for examples).

[2] Sketch Engine, for example, defines a *token* as the smallest unit that a corpus consists of. Tokens can be split into *words* and *non-words*. The former refers to tokens which begin with a letter of the alphabet, while the latter refers to tokens which do not start with a letter of the alphabet (e.g., a punctuation mark or a digit).

[3] (http://ucrel.lancs.ac.uk/claws).

[4] (https://ucrel.lancs.ac.uk/usas).

https://doi.org/10.1017/9781009168144 Published online by Cambridge University Press

So what does all this have to do with the exploration of discourse? There are two things that we can do based on frequency information alone: (i) compile a wordlist, which orders linguistic units (lexical, grammatical, or semantic) alphabetically or by frequency; or (ii) run searches for individual linguistic units, and compare their frequencies across corpora or parts of corpora (referred to as *subcorpora*). These word lists can be very useful indeed, especially if researchers know where to direct their search. For example, making use of the tagged data, we might decide to focus on content words (e.g., nouns or verbs) rather than grammatical words (e.g., determiners, prepositions, pronouns), because the former are more semantically loaded and thus give us an idea of what themes can be found within a corpus – even before we begin looking into the surrounding co-text and the social context.

To demonstrate this process, we will draw upon two worked examples where we examine the wordlists of two specialised corpora (see also Mautner, 2022: 254–5). The first wordlist is from the UK Supreme Court (UKSC) corpus, which consists of all judgements between 2009 and 2018 that contain at least one dissenting opinion (129 judgements; 3,376,434 tokens). The top ten most frequent nouns in the corpus can be found in Table 1. While it is probably unsurprising that these nouns are mainly legal terms, it may be interesting to note that half of them refer to other texts (*case*, *section*, *para* [paragraph], *act*, and *article*). Immediately, this sparks further questions around Supreme Court judgements' highly intertextual nature, as they contain numerous cross-references to other texts, ranging from the judgements being appealed against to arguments being put forward in the public hearing and the precedents being invoked (Mautner, 2022: 254).

Rank	Noun	Raw frequency
1	court	15,695
2	[any number]	15,509
3	case	13,170
4	section	9,617
5	para	8,854
6	lord	8,144
7	law	7,893
8	act	7,460
9	right	6,690
10	article	6,644

Table 1 The top ten most frequent nouns in the UKSC corpus

Rank	Noun	Raw frequency
1	[any number]	111,760
2	organization	75,640
3	firm	44,558
4	group	35,517
5	study	35,010
6	research	31,778
7	model	29,932
8	effect	28,953
9	work	27,461
10	variable	27,358

Table 2 The top ten most frequent nouns in the ASQ corpus

The second example is from the *Administrative Science Quarterly* (ASQ) corpus, which consists of every article and book review published in the journal from its inception in 1956 through to the end of 2018 (3,547 articles; 19,470,470 tokens) (Mautner and Learmonth, 2020). Table 2 shows the top ten most frequent nouns in the corpus.

In a similar vein to our previous example, it may again be unsurprising that the top ten nouns of the ASQ corpus are related to academic and organisational discourse, but three of them are self-referential (i.e., *study*, *research*, and *model*) in that they are referring back to something in the same or other academic papers (Mautner, 2022: 254). We know this to be characteristic of academic writing, in that (theoretically speaking, at least) each new paper further develops a field of work that has come before it. As a first step towards building up a discursive profile of the two corpora, and before using any other corpus linguistic tools this is already a useful exercise. It gives us an indication of what the corpora are about, and where we might wish to direct future searches and analyses.

We have just shown how individual words can give an insight into the discourse manifested by the corpus, but it might be equally fruitful to explore multi-word units. These are called *n*-grams, where *n* refers to the number of words found within the unit. It is generally the case that the higher the *n*, the more distinctive that phrase is likely to be of that particular corpus – a 2-gram such as *in the* is likely to be found across all genres, whereas a 6-gram such as *decision of the Court of Appeal* is much more specific to our particular corpus of Supreme Court judgements. Sketch Engine gives us the option to determine an *n*-gram length, and then explore the most frequent *n*-grams within a particular corpus. Starting with a simple list of highly conventionalised and formulaic language, before delving deeper into the context, can again go some way towards telling us about the type of discourse enacted by the corpus.

Useful though they are, wordlists may easily mislead us. We may assume, for example, that a wordlist provides the quantitative evidence for salient themes within a corpus, and it is therefore immediately meaningful to us as discourse analysts. This may well be the case, of course, but it is also worth asking what this salience is measured in relation to. If, over the course of an investigation, we are making claims about which items are central to a corpus, it might be worth checking the frequency of those items in a general reference corpus of the English language. Intuition may tell us that a particular word is salient, but in reality that may not be the case.

When checking such claims in another corpus, it is useful to compare the *relative* (or *normalised*) frequency, rather than the *raw* frequency. This is because corpora can differ wildly in size, and we need to ensure that we are comparing like with like. We would be in trouble, for example, if we compared the raw frequencies of the noun *court* across both the UKSC corpus and the British National Corpus 1994 (BNC1994) because the former is only around 3 per cent of the size of the latter. Within CL, it is good practice to calculate the relative frequency according to the size of the corpus: if the corpus size is under 100,000, we would normalise per 10,000 words; if it was under 1 million, we would normalise it per 100,000 words; and so on. This means that the relative frequencies are closer to the actual (raw) frequencies. To calculate the relative frequency, we take the raw frequency of the word under investigation, divide it by the total number of words in the corpus, and then multiply the result by either 10,000 or 100,000 or 1 million, depending on the size of the corpus. What we find, then, is that the noun *court* occurs 4,648 times per million tokens in the UKSC corpus, whereas in the BNC1994 it only occurs 247 times per million tokens: a result that we could perhaps have expected, given our knowledge of the two corpora, but checking intuition in this way is highly important.

One further issue to be aware of when interpreting frequency information is that words may be clustered and unevenly distributed across the corpus (i.e., showing an uneven *dispersion*). This may occur when a small number of texts within the corpus use a certain word or phrase so often that it moves to a high position on the wordlist, or is considered to be key as the result of a keyword analysis. There is a danger that when viewing such a list, we implicitly assume it is representative of the corpus, when in reality it may stem from a smaller number of texts. Such clustering may be interesting from a discourse perspective, as it could give us insights into unique language usage, or patterns that are only used within particular discourse communities. This was apparent in

https://doi.org/10.1017/9781009168144 Published online by Cambridge University Press

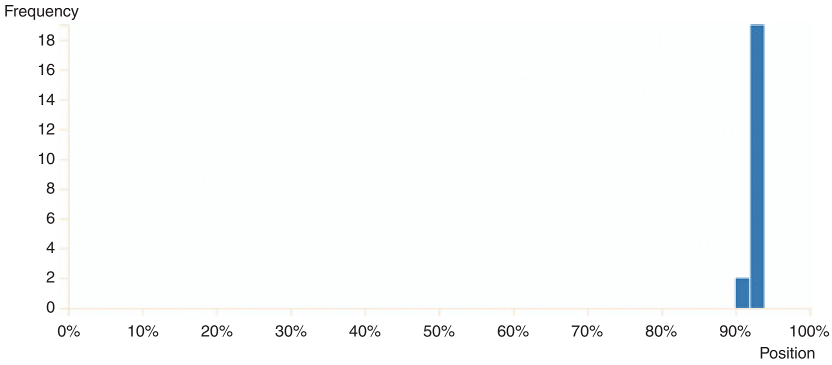


Figure 1 Dispersion of *focal leader* within the ASQ corpus using Sketch Engine

Mautner and Learmonth (2020), who explored social actor representation and neoliberal rhetoric in the ASQ corpus. While examining the collocates of *manager* and *leader*, the adjective–noun combinations *focal manager* and *focal leader* appeared on collocate lists with a high logDice score (more information on different collocation statistics in Section 4.3). Originally, these expressions were thought to be popular within the discourse community of scholars working within organisation studies; in reality, they were used across just three texts, and were seemingly coined by the authors of those articles and not picked up any further (indeed, two of the three papers were authored by the same scholar). The benefit of using corpus analysis programs is that we can take a closer look at such clusters to disentangle what is happening in, and perhaps what is special about, the texts concerned. It provides a safety net to ensure we are not being misled by frequency alone.

This example presents a very simple way of exploring dispersion. Corpora that have been compiled with an appropriate amount of useful metadata should enable us to nip such issues in the bud using similar checks and balances. Both AntConc (Anthony, 2022) and Sketch Engine similarly provide an easy way to visualise word occurrence across a corpus. In Figures 1 and 2, we demonstrate Sketch Engine's *distribution of hits* function, which allows the user to see where particular words occur within files. In Figure 1, we can see the phrase *focal leader* appearing in just two texts, with the distribution represented in a bar chart – one text contains the phrase twice, and another text contains the phrase nineteen times. In Figure 2, we can see how the phrase *senior manager* is much more widely distributed, and we can alter the granularity of the chart to find out which texts use the phrase most frequently.

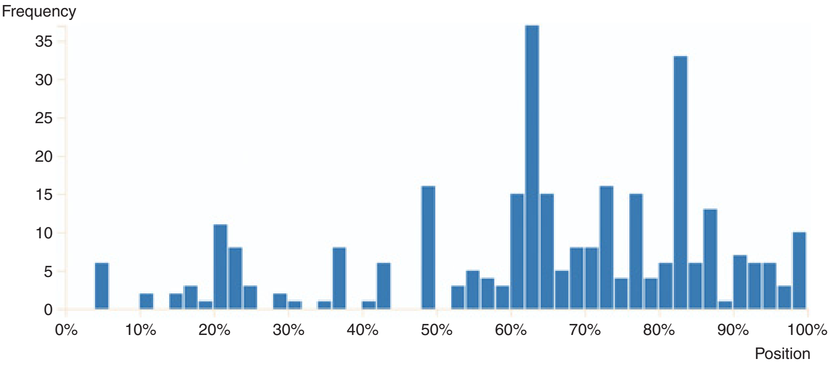


Figure 2 Dispersion of *senior manager* within the ASQ corpus using Sketch Engine

For every part of the corpus analysis process, including dispersion metrics, it is important to consider the underlying file structure within the corpus. If you want to be able to make statements about individual files, they need to remain as identifiable, self-contained units. These units could be, for example, individual newspaper articles or conversations, or they could be everything said by a particular speaker. Egbert and Schnur (2018) argue that the text should be both the sampling unit (in corpus construction) and the observational unit when it comes to the analysis. Being aware of exactly how the corpus is constructed from the ground up is key to accurate interpretation and understanding.

More advanced dispersion metrics are increasingly being incorporated into corpus analysis programs and being accounted for in analyses (for overviews, see Gries, 2008; Brezina and Meyerhoff, 2014; Biber et al., 2016; Brezina, 2018). CQPweb recently introduced a dispersion plot which visualises how the frequency of a particular lexical item maps onto each file within the corpus (Gomide, 2020). The tool provides dispersion metrics (such as DPnorm (Gries, 2008) and Juilland's D (Juilland, Brodin and Davidovitch, 1970)) to allow different words to be easily compared. Depending on the individual research project, it may also be appropriate to employ advanced statistics, such as so-called mixed-effects models, which appear to be gaining traction within the field of CL more generally. Mixed-effects models offer a powerful and robust method to assess variation within a corpus (Gries, 2015; Brezina, 2018; Gillings, 2021).

To sum up, then, frequency and dispersion have an important role in CADS. Frequency is at the heart of the endeavour, allowing us to identify the most prominent (or neglected) discourses within a corpus. Dispersion, by extension, is one of the checks and balances available to us, allowing us to explore the extent to which those discourses are spread throughout the whole corpus, or whether they are restricted to a much smaller sample. Taking both techniques into account when carrying out our analyses allow us greater confidence in our findings and subsequent interpretation.

Key takeaways and things to watch out for:

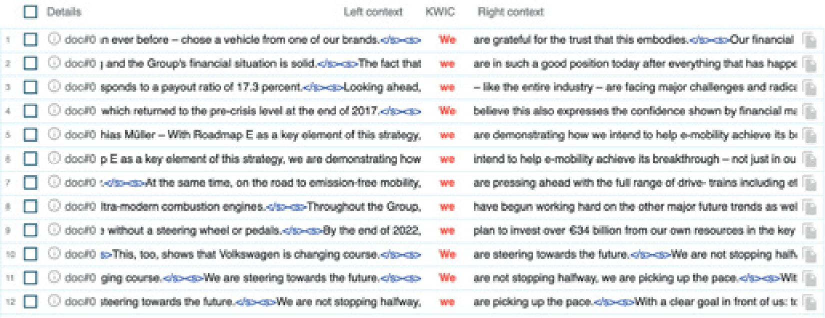
- Corpora can be tagged for grammatical part-of-speech and semantic domains.
- Wordlists (or indeed, part-of-speech or semantic lists) can give us some indication as to the discursive representations prevalent in a corpus.
- Exactly what constitutes a ‘word’ is a source of debate. Different researchers and different tools tokenise corpora in different ways.
- Using relative frequency, rather than raw frequency, allows us to make comparisons across corpora of different sizes.
- It is important to check dispersion to determine whether words are clustered in specific texts or whether they are spread throughout the whole corpus.

4.2 Concordance Analysis

Concordance lines are a core element of CL; something which, at least visually, makes a corpus linguistic analysis stand out from other computer-assisted work (Sinclair 1991; Evison, 2010; Tribble, 2010; Hunston, 2022). Put simply, concordance lines allow the user to view a search term or phrase within its linguistic environment. The search term (i.e., the ‘node’ or ‘node word’) tends to be centred down the middle of the screen, with its co-text stretching off to the left and right (hence KWIC, ‘Key Word in Context’), thereby allowing the user to read through examples as efficiently as possible. Figure 3, a screenshot from Sketch Engine, demonstrates this. Here, we can see (almost) at a glance, that the majority of occurrences of *we* (all except line #3) are used with verbs describing broadly positive activities or emotions: a linguistic choice in keeping with the promotional nature of the text.

Concordance lines can often be sorted in various ways (randomised or alphabetised), thinned (to work with a more manageable sample), and, crucially for CADS, expanded to view more co-text – perhaps up to the level of the sentence, paragraph, or even the complete text. Some programs, such as CQPweb and Sketch Engine, also allow the user to assign a category to individual concordance lines. Figure 4 shows what this feature looks like on CQPweb. The so-called categorisation query can be given a name, as can the individual categories. After setting up the query, a drop-down menu appears next to each concordance line, allowing the researcher

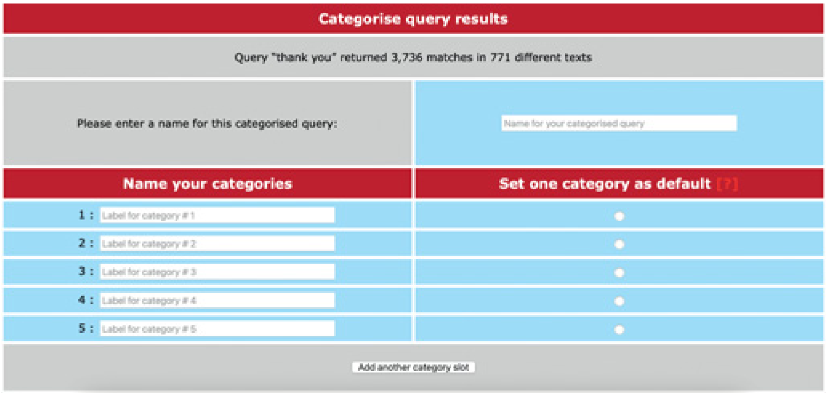
https://doi.org/10.1017/9781009168144 Published online by Cambridge University Press



	□	Details	Left context	KWIC	Right context
1	□	ⓘ doc#0	n ever before – chose a vehicle from one of our brands.</s><s>	We	are grateful for the trust that this embodies.</s><s>Our financial
2	□	ⓘ doc#0	j and the Group's financial situation is solid.</s><s>The fact that	we	are in such a good position today after everything that has happe
3	□	ⓘ doc#0	sponds to a payout ratio of 17.3 percent.</s><s>Looking ahead,	we	– like the entire industry – are facing major challenges and radic
4	□	ⓘ doc#0	which returned to the pre-crisis level at the end of 2017.</s><s>	We	believe this also expresses the confidence shown by financial m
5	□	ⓘ doc#0	hias Müller – With Roadmap E as a key element of this strategy,	we	are demonstrating how we intend to help e-mobility achieve its b
6	□	ⓘ doc#0	p E as a key element of this strategy, we are demonstrating how	we	intend to help e-mobility achieve its breakthrough – not just in ou
7	□	ⓘ doc#0	:</s><s>At the same time, on the road to emission-free mobility,	we	are pressing ahead with the full range of drive- trains including el
8	□	ⓘ doc#0	ltra-modern combustion engines.</s><s>Throughout the Group,	we	have begun working hard on the other major future trends as wel
9	□	ⓘ doc#0	s without a steering wheel or pedals.</s><s>By the end of 2022,	we	plan to invest over €34 billion from our own resources in the key
10	□	ⓘ doc#0	s>This, too, shows that Volkswagen is changing course.</s><s>	We	are steering towards the future.</s><s>We are not stopping half
11	□	ⓘ doc#0	ging course.</s><s>We are steering towards the future.</s><s>	We	are not stopping halfway, we are picking up the pace.</s><s>Wit
12	□	ⓘ doc#0	steering towards the future.</s><s>We are not stopping halfway,	we	are picking up the pace.</s><s>With a clear goal in front of us: t

Figure 3 A random selection of lines from the concordance of *we*, occurring in the *Letter to our Shareholders* included in the *Volkswagen Annual Report 2017*.

to systematically go through and code the lines as necessary. It is worth noting, though, that even if your preferred corpus analysis program does not have this



Categorise query results

Query "thank you" returned 3,736 matches in 771 different texts

Please enter a name for this categorised query: Name for your categorised query

Name your categories	Set one category as default [?]
1 : Label for category # 1	○
2 : Label for category # 2	○
3 : Label for category # 3	○
4 : Label for category # 4	○
5 : Label for category # 5	○

Figure 4 CQPweb's concordance line categorisation function

feature built-in, it is commonplace to download concordance lines by exporting them into Excel, and then categorising them there.

Being able to see so many examples of a word in context, at a glance, means that these programs are a powerful heuristic tool. They allow the user to detect patterns of usage, and through doing so, build up discursive profiles of particular items. Concordance lines make it possible to explore – to paraphrase Firth (1957: 11) – the company that a word keeps.

While in this Element we discuss concordance lines as they are used by CADS researchers, discourse analysis is by no means the only field that has benefitted from organising data in such a way. Grammarians and syntacticians, for example, consult concordance lines to determine how particular words 'behave' in different linguistic environments, and lexicographers consult concordance lines to determine a word's shades of meaning and identify examples suitable for a dictionary. In the field of language learning and teaching, there is a subdiscipline known as *data-driven learning*, which advocates the use of concordance lines in the classroom; that is, using them as a pedagogical tool for learners to see how authentic language is used in context (see Pérez-Paredes and Mark (2021) for an overview). Indeed, many fields, across linguistics and beyond, have seen value in presenting textual data in the concordance format.

A CADS perspective differs from these approaches in one crucial respect: discourse is the focus of analysis, and corpus assistance helps us to link large-scale social phenomena with linguistic choices at the micro level. Indeed, CADS scholars are less concerned with a word's syntactic position for its own sake, and are more interested in whether that syntactic position says anything about discursive representation (whether, for example, a particular group of people is predominantly portrayed as active or passive); they are less concerned with the range of meanings a word has and are more interested in how that meaning is built up and reinforced by speakers in a particular discourse context. Thus, CADS work focuses on linguistic signs less for what they *are* and how they are related to one another, and more on what signs *do* and how they are related to the extra-linguistic world.

Due to this interest in the social function of language, we also need to look beyond the concordance line. Here, 'beyond' means two things: on the one hand it means reading and interpreting not just the line itself, but an expanded stretch of co-text before and after that line (a feature that all of the most frequently used corpus analysis programs offer with one mouse-click). On the other hand, we need to go beyond the concordance line in the sense of relating it to the wider social context which shapes the corpus and is shaped by it. Such a contextually sensitive perspective is, of course, the 'home turf' of traditional non-computer-based discourse studies (DS); a perspective we do not have to abandon entirely, precisely because we can expand concordance lines if and when necessary. In effect, CADS scholarship uses the concordance as a window through which complete texts can be accessed.

So how does one actually conduct a concordance analysis? In Figure 5, we present four prototypical approaches. This figure is by no means intended to be prescriptive. Instead, it is a distillation of current scholarship, designed to explain how concordance analysis differs along two main axes: the extent to which it is bottom-up or top-down (i.e., led by data or theory), and the extent to

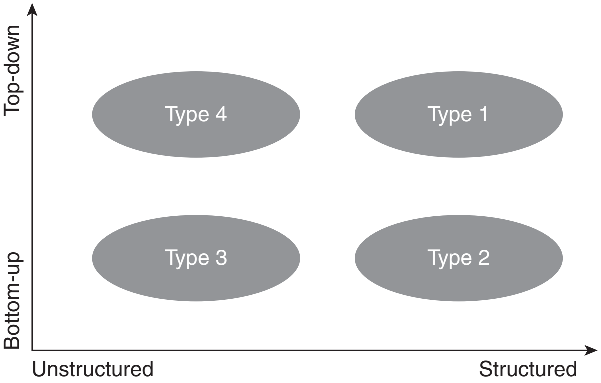


Figure 5 Four common types of concordance analysis (Gillings and Mautner, forthcoming)

which it is structured or unstructured (i.e., systematic or exploratory). As proposed in Gillings and Mautner (forthcoming), we outline the four types of concordance analysis as thus:

- **Type 1:** a structured top-down process whereby the researcher applies an *a priori* framework or set of categories to a number of concordance lines (e.g., Culpeper and Gillings (2018) coding for politeness in the BNC1994/2014, and Lutzky (2021b) exploring the pragmatic functions of *sorry* in customer service interactions on Twitter);
- **Type 2:** a structured bottom-up process whereby the researcher assigns categories to the concordance lines, but these come organically from the corpus rather than being imposed on it (e.g., Kopf (2019) exploring the ways that content policies are enforced on Wikipedia, and Zottola et al. (2021) identifying coping strategies that patients use in autobiographical narratives while waiting for assessment at a transgender health clinic);
- **Type 3:** an unstructured bottom-up process whereby the researcher eyeballs the concordance lines and lets that qualitative holistic judgement form the basis of analysis (e.g., McEnery, Baker and Dayrell (2019) identifying previously unrecorded droughts in nineteenth-century Britain, and Levon (2016) exploring the extent to which users on a question-and-answer forum use their replies as an opportunity for stance-taking);
- **Type 4:** an unstructured top-down process whereby the researcher identifies concordance lines which match categories proven to be relevant in other datasets (e.g., Archer and Gillings (2020) identifying potential indicators of deception in Shakespeare's plays, and Appleton (2021) exploring how the unification of Germany is discussed in Hansard).

https://doi.org/10.1017/9781009168144 Published online by Cambridge University Press

Types 1 and 2 both call for the researcher to sift through each and every concordance line within a sample, but they differ with regard to whether categorisation is something that is applied to, or extracted from, the data. Types 3 and 4 are similar in that they are both on the more exploratory end of the scale, but again differ with regard to whether the corpus is used to form a holistic judgement, or whether it is a means to identify relevant concordance lines. However, it is completely possible that some form of semi-structured analysis could take place, or even one that starts top-down and morphs into something that is more bottom-up as the analysis progresses. This middle ground is often useful when there is a high number of concordance lines; the researcher may decide to categorise a smaller number of lines, calculate the percentage for each category, and then scale up those frequencies to the full set. In actual research practice, the four types are not entirely discrete but are best thought of as occupying potentially overlapping spaces.

To briefly offer a worked example, we turn to Mautner (2005), which uses concordance analysis to explore how the phrase *a/the entrepreneurial university/ies* was used in WebCorp (a site which allows one to explore the web, in real time, as a corpus; www.webcorp.org.uk/live). Serving as an entry point to the discourse around university branding, Mautner (2005) downloaded 121 concordance lines and categorised them according to whether they had a positive or negative semantic prosody. *Semantic prosody* is a term used by Louw (1993: 157) to refer to ‘the consistent aura of meaning with which a form is imbued by its collocates’. Mautner (2005) found that 63 out of 121 usages had a positive semantic prosody, 25 had a negative one, and the remaining 33 were unclassified because the concordance line did not offer any evaluative lexis to aid deduction. Concordance lines with a positive semantic prosody found the query phrase alongside adjectives such as *strong*, *modern*, *dynamic*, and *top*; adverbs such as *highly*, *distinctively*, and *truly*; predicates such as *renowned for*, *is proud of being*, and *showcases us as*; and phrases such as *active and competitive*, *of high quality*, and *one of the most self-sufficient in the country*. Concordance lines with a negative semantic prosody, on the other hand, found *a/the entrepreneurial university/ies* alongside emotive language such as *pernicious ideologies* and *capitalist regime*, and negatively loaded lexis such as *challenge to*, *limitations of*, and *alternatives to*. This co-text can be explored and categorised, and the source text can be explored further to ensure the linguistic constructs are interpreted within the wider discourse and context. What is clear, then, is that concordance analysis is not only a good initial way-in to the data but also a fruitful way to achieve a more nuanced picture of the discourse concerned. Furthermore, concordances help us see the traces of more

https://doi.org/10.1017/9781009168144 Published online by Cambridge University Press

'minor' discourses, which may still be highly relevant, but not as salient as their 'major' counterparts that define the genre (Baker, 2015).

Key works on conducting concordance analysis include Sinclair (1999, 2003), Hunston (2002), and Baker (2006). Baker (2006: 92–3) specifically applies the procedure to discourse analysis and incorporates other corpus linguistic methods into the framework, before conducting the concordance analysis. Among other things, he suggests noting 'rare or non-existent cases based on your own intuitions' (e.g., discourses which stand out as being notably absent; see also Schröter and Taylor, 2018) and to investigate patterns in the corpus with the help of collocates and comparative data from reference corpora (Baker, 2006: 92–3). Here, Baker stresses that the onus is on the researcher, not the tool, to interpret the saliency of themes on the basis of their own knowledge and background. In addition, when combining concordance analysis with other corpus linguistic techniques, it is the researcher's responsibility to decide at which point concordance lines become most useful. In practical terms, then, a concordance analysis could stand alone, but it could also be the end point after other techniques have been employed. Critically speaking, 'a concordance analysis is therefore only as good as its analyst' (Baker, 2006: 89).

It is this kind of critical reflection on concordance analysis that we think deserves further attention. After all, on their most basic level, concordance lines are simply the output of a search tool; they do not speak for themselves but need to be interpreted. In a recent research project, Gillings and Mautner (forthcoming) identified eight interpretability issues that CADS scholars are likely to encounter when reading through concordance lines. These barriers to successful interpretation, in order of frequency in the sample used for the investigation, are as follows: use of technical terms and jargon; referring expressions pointing to a lexical item outside the available co-text (so that we do not know what a particular pronoun refers to, for example); lines that are unrelated to the research question; use of acronyms and initialisms; instances where the co-text is so unspecific that it is impossible to deduce any meaning at all; an unclear quotation source attribution (so that we cannot tell who said what); non-standard syntax; and noise in the corpus (Gillings and Mautner, forthcoming). Some of these are potentially highly problematic for CADS work. For example, if a concordance line displays a particularly long quotation, the constraints of the concordancing software may mean that we are unable to see the quotation marks on each side of the quote. We may therefore end up reading a quotation without realising that it is a quotation, which can lead to a whole host of issues – after all, whose voice we hear is of central concern in discourse analysis. As expected, sometimes these issues can be fixed easily by simply expanding the concordance line to view more co-text, but at other times this may be more

https://doi.org/10.1017/9781009168144 Published online by Cambridge University Press

problematic than one might initially have imagined. Likewise, some issues may only be solved by reading the whole text, or reading related texts not found in the corpus at all.

Taking account of these issues, Gillings and Mautner (forthcoming) also give advice on concordance line interpretation, informed by the principles of transparency and self-reflective critique, as well as by earlier work on concordance analysis such as Sinclair (2003), Hunston (2002), and Baker (2006). Among other things, Gillings and Mautner recommend filtering concordance lines in waves and carefully documenting the range of reasons why concordance lines were removed from the sample; distinguishing between concordance lines that are genuinely uninterpretable and those that are simply irrelevant to the research question; and finally, guarding against overinterpretation of the data by refining the research questions and search procedure. The latter is not only in line with Sinclair's (2003) recommendations to revise hypotheses when presented with new patterns in the concordance but this process of refinement is also a general principle relevant at every stage of the research process. However overwhelming our desire to produce findings, we must resist the temptation of claiming that there are patterns in the data without offering compelling evidence.

Key takeaways and things to watch out for:

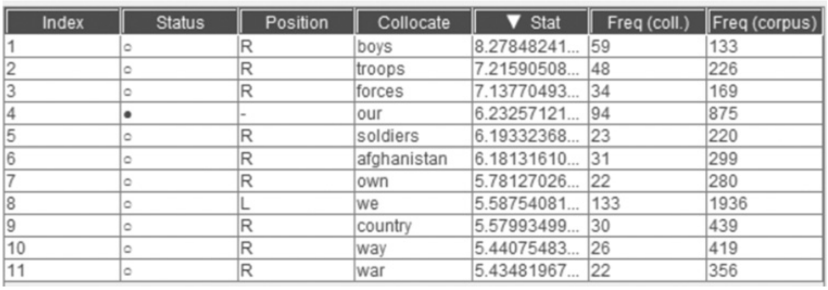
- Concordance analysis can take various forms, differing according to whether it is structured or unstructured, and whether it is bottom-up or top-down.
- *Co-text* refers to the immediate language around the node word, whereas *context* refers to the wider social environment.
- Identifying positive and negative semantic prosodies can tell us a lot about how discourses are constructed. Minor linguistic details can play a major part in this.
- Concordance lines are a product of the corpus analysis program, not a unit of meaning in and of themselves.
- Issues with interpretability mean that care must be taken to read concordance lines and the surrounding co-text carefully, rather than rashly making assumptions about what is found at first glance.

4.3 Collocation Analysis

Collocates are words which frequently co-occur, more often than would otherwise be expected by chance alone. They are useful for discourse analysis because they bestow additional meanings on words which can

https://doi.org/10.1017/9781009168144 Published online by Cambridge University Press

help to indicate the author's viewpoint and value judgements shared by speech communities. The words may sometimes occur in a fixed position – for example, within idiomatic phrases like *bits and bobs*, or they may occur within the same range but within different positions (e.g., *tell a story*, *a story to tell*). The chosen window within which two words co-occur will impact on both the type and number of collocates that are obtained. For example, restricting the collocational span to one word (i.e., the word that occurs immediately before or after the search word) is likely to result in fewer collocates and will favour certain types of relationships, such as adjectives followed by nouns. On the other hand, using a span which takes into account five or even ten words on either side of the word being investigated is likely to produce more collocates, of higher frequency, and a wider range of grammatical types. That said, the wider range may also yield some collocates where the relationship between the two words is more tenuous. It is worth carrying out experiments on the corpus you are working with to identify which span is likely to provide a representative and useful set of collocates based on the research aims. Most corpus tools allow collocates to be calculated via an enquiry on a particular word, resulting in a list usually ordered by the score associated with the statistic used. This is demonstrated in Figure 6.



Index	Status	Position	Collocate	▼ Stat	Freq (coll.)	Freq (corpus)
1	○	R	boys	8.27848241...	59	133
2	○	R	troops	7.21590508...	48	226
3	○	R	forces	7.13770493...	34	169
4	●	-	our	6.23257121...	94	875
5	○	R	soldiers	6.19332368...	23	220
6	○	R	afghanistan	6.18131610...	31	299
7	○	R	own	5.78127026...	22	280
8	○	L	we	5.58754081...	133	1936
9	○	R	country	5.57993499...	30	439
10	○	R	way	5.44075483...	26	419
11	○	R	war	5.43481967...	22	356

Figure 6 Collocates of the word *our* in a corpus of news articles about Muslims from *The Sun* newspaper. Analysis carried out using #LancsBox, ranked by the Mutual Information (MI) statistic. ('Freq (coll.)' tells us how frequently *boys*, for example, co-occurs with *our*, whereas 'Freq (corpus)' tells us how frequently *boys* occurs within the whole corpus.)

The most basic way of identifying collocates is simply to count the number of times one word co-occurs with another word in a corpus. A potential issue with this approach is that it is likely to privilege high-frequency grammatical words like *the* or *to*. This is why these are sometimes removed from the list after computation in order to focus instead on high-frequency noun, verb, and adjective collocates. Another option is to use a collocational statistic. Some, like MI, focus on the strength of the relationship between two words, so will favour cases where the words often appear together as opposed to being apart. This often gives prominence to low frequency pairs like *bits* and *bobs*. Other statistical tests, like log-likelihood tests, focus on the amount of evidence that a relationship exists between two words. They will thus often favour higher-frequency words (such as *give* and *him*), although such relationships are less likely to be exclusive as *him* can often occur some distance from *give*, and vice versa. More recent measures like logDice have attempted to take both types of relationship into account, offering a kind of compromise. An accessible overview of various statistical concepts and techniques is provided by Brezina (2018).

The uniting factor, between these techniques, is that they all try to identify pairs of words that co-occur in the corpus with a frequency that is greater than chance. It is assumed that whenever that is the case, it tells us something about a recurring discursive construction (such as a particular group of people always being described with a particular adjective). It is assumed further that such a finding has predictive power. In other words, if we build a corpus along similar lines, then we would expect to find the same patterns of co-occurrence.

It is up to the analyst to decide how high the logDice, MI, or log-likelihood score needs to be in order for a word to count as a collocate. Focusing on MI in particular, it has been suggested that the value would need to be at least 6 in order to lead to 'collocational priming', whereby encountering one word triggers an association with another (Durrant and Doherty, 2010). Depending on the size of the corpus and the frequency of the node word, imposing such cut-offs may result in a tiny number or dozens of collocates, so sometimes analysts decide to consider only a set number of collocates from the top of each list. If you are dealing with, say, the first ten collocates, each can be explored in detail. If on the other hand, you want to look at the first fifity collocates, it might be better to group them into categories, and then examine one or two words from each category in more detail. In any case, it is important to be clear about which methods, types of categorisation and cut-off points were used, including some information about the types of collocates which those parameters are likely to produce, and why this was particularly useful for your analysis.

Many words may appear at first glance to have fairly neutral meanings but as a result of repeatedly occurring in certain contexts, they can gather negative or positive associations which can influence readers or hearers in a certain

https://doi.org/10.1017/9781009168144 Published online by Cambridge University Press

direction. This relates to the concept of semantic prosody (Louw, 1993; Partington, 2004), as discussed in the previous subsection. For example, in a corpus of propaganda texts which advocated violent jihad, the word *America* frequently collocated as the agent of the verbs *invade*, *kill*, *attack*, *perpetrate*, *launch*, *desecrate*, *inflict*, and *violate* (Baker, Vessey and McEnery, 2021). This would indicate a negative prosody around *America* in this corpus, particularly as concordance analysis showed that these verbs normally occurred when America was the agent (i.e., the social actor carrying out these activities). While this is a very specific prosody which we would not be as likely to find in other corpora, some words have more flexible prosodies, appearing in a wider range of text types. To use one of Stubbs' (2001) examples, the verb *cause* means 'make something happen', and does not appear to have a positive or negative prosody. However, in general reference corpora it collocates with words like *damage*, *death*, *disease*, *pain*, and *trouble*, indicating that it has a negative prosody which can be taken advantage of by authors to signify a certain stance, even when used in seemingly positive contexts. For instance, the phrase *it caused amusement* could be used to indicate that the author disapproves of people finding something amusing. Thus, a collocational analysis can be useful in spotting hidden or subtle meanings which can signal the author's stance.

Furthermore, a collocational analysis can be used to show differences between related words or concepts. Baker (2014) examined collocates of the words *boy* and *girl* (along with their plurals), finding that *girl(s)* tended to collocate with words that expressed emotional states like *smile*, *suffer*, *love*, and *want* as well as adjectives like *shy*, *fond*, *anxious*, *unhappy*, *eager*, and *desperate*. *Boy(s)* on the other hand collocated with words involving physical actions or processes like *play*, *fall*, and *die*. *Girl(s)* was more likely to collocate with words which represented them as victims like *rape*, *abduct*, *murder*, and *assault*, whereas boys were more likely to attract words relating to them as either behaving poorly or well (e.g., *naughty*, *bad*, *golden*, *genius*, *credit*).

Some corpus tools also allow more sophisticated analyses of collocates. For example, Sketch Engine's WordSketch feature groups collocates according to their grammatical relationships, so if our node is a noun like *mother*, the verbs which position *mother* as a grammatical agent (i.e., the 'doer') will appear in one list, while other verbs which position *mother* as a grammatical patient (i.e., the 'recipient') will appear in another list. Adjective modifiers appear in a third list and so on. If we had used software that didn't have this feature, we could have categorised the collocates by hand into the same grammatical groups, but it would have taken more time and required a great deal of concordance analyses. Sketch Engine can also compare two words (like *boy* and *girl*), identifying the extent to which they have shared and unique collocates, which takes a lot of the manual work out of a comparative analysis. Sketch Engine's categorisation is obviously much faster than could be done manually. We should bear in mind, though, that automatic categorisations are rarely 100 per cent accurate; occasionally, some words will need to be excluded or moved to a different group.

Another tool, #LancsBox, allows analysts to create collocational networks. It produces a visual representation of a set of collocates and shows how they are linked to one another. Analysing a corpus of articles from the British newspaper *The Sun*, Baker (2016) describes how this tool can be useful from a discourse analysis perspective. Figure 7, taken from that paper, shows a network indicating which expressions are used to refer to members of the British army who were fighting in Afghanistan: for example, *British soldiers*, *British troops*, *our soldiers*, *our troops*, *our forces*, *our boys*. In terms of their dictionary meaning, these naming strategies are (near-)synonyms, but obviously their connotations are very different; and choosing one over the other has a significant impact on how events are framed. The network opened up a line of analysis which involved examining the contexts in which the more formal (*British soldiers*) or affectionate (e.g., *our boys*) terms were used.

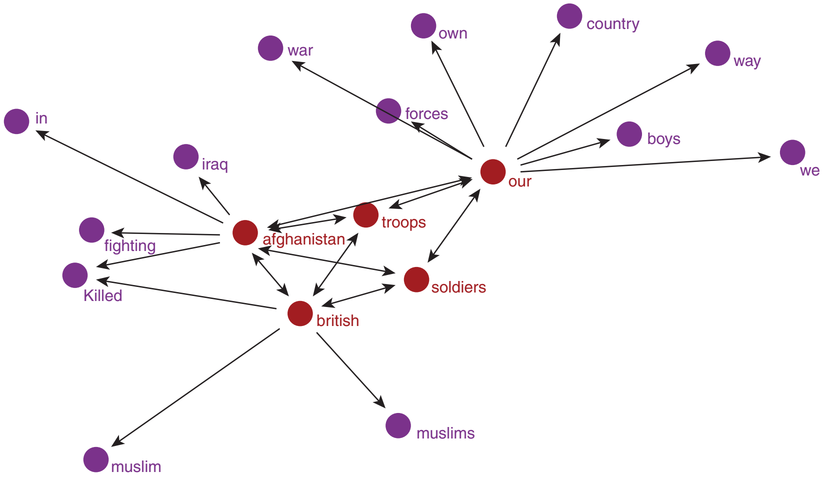


Figure 7 Collocational network from a corpus of British newspaper articles (Baker, 2016: 145)

The value of collocational data is beyond dispute; yet it is important not to overinterpret it. A table or visual network of collocates doesn't tell us much about how two words actually relate to one another. This can only really be ascertained through a more detailed concordance analysis. In cases where there are hundreds of concordance lines, focusing on a random sample of just 100 or so is likely enough to enable a generalisation to be made about how the words in question work together. However, if a word produces very few collocates and/or all the collocates occur with very low frequencies, then a collocational analysis might not be the best approach to take. In those cases, too, a fine-grained concordance analysis of the node word is the answer.

When dealing with large corpora or very frequent words, it can also be useful to focus on a smaller number of possible collocation patterns from the outset. For example, in Baker and Baker (2019), collocates of the words *masculinity* and *femininity* were examined in a corpus of newspaper articles. Collocates that occurred in three different grammatical patterns were considered: [NOUN] *of masculinity/femininity* (e.g., *expressions of femininity*), [VERB] [PRONOUN] *masculinity/femininity* (e.g., *strutting his masculinity*), and [ADJECTIVE] *masculinity/femininity* (e.g., *gentle femininity*). This allowed for a closer focus on noun, verb, and adjective collocates that occurred in particular patterns, rather than all possible contexts. The analysis found, for example, that masculinity was often personified as a strong body (e.g., *pumped*, *rugged*, *muscular*, *sweating*, or *toned masculinity*), while femininity was described as soft and passive (e.g., *delicate*, *sweet*, *demure*, or *dainty femininity*).

When working with collocations, it is worth taking into account the distribution of frequency data across different texts (i.e., its dispersion). If two words are found to collocate strongly, but this relationship only occurs within one or two texts in a corpus that is made up of hundreds of texts, then we are not really seeing a generalisable pattern but instead something which is likely to be idiosyncratic. Some analysts impose distribution cut-offs when considering collocates; for example, by specifying that a collocational relationship needs to appear in at least five texts or 1 per cent of the texts in the corpus. However, the less well-distributed collocational relationships might still be worth commenting on, because they might point to a minority or rare discourse that is socially significant.

Key takeaways and things to watch out for:

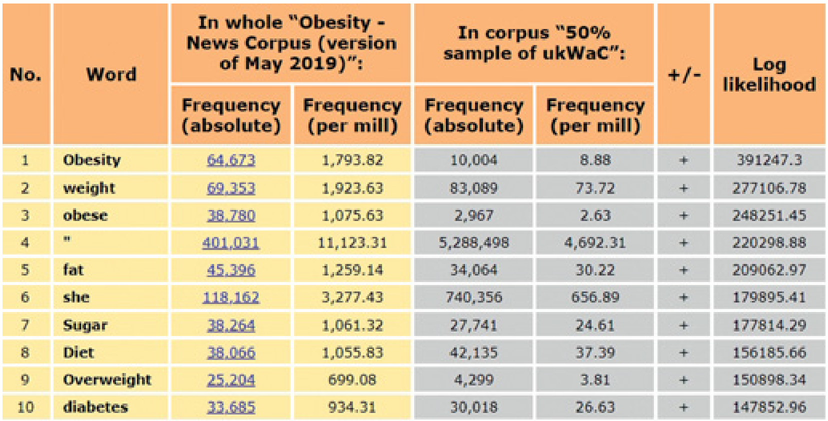
- Collocation analysis can identify repeated relationships between words. This is useful for identifying ways in which authors indicate stance, either explicitly or subtly, along with offering an insight into widely accepted ways of representing concepts.

- There are different ways of calculating collocation, although none of them are perfect. Analysts should be aware of the types of collocates that their chosen technique is likely to favour.
- Both the collocate categorisation tool offered by Sketch Engine ('WordSketch') and the collocational network tool offered by #LancsBox can provide nuanced understandings of collocational patterns. However, such tools do not always provide further insights, and a manual analysis may be more fruitful. They should not be employed just because they exist.
- It is often infeasible to examine every collocate, and doing so would result in a repetitive analysis. It can be a good idea to narrow the focus to a selection of representative collocates, each of which may tell us something different.
- The study of collocates works best when coupled with concordance analyses, and it is also worth taking into account the distribution of collocational patterns across a corpus.

4.4 Keyword Analysis

A keyword is a word which occurs statistically significantly in one corpus when compared against a second corpus. The two corpora under comparison are usually related in some way. For example, if we want to analyse the discourse of obesity in a study corpus of British newspaper articles (as in Brookes and Baker, 2021), we might compare it to a reference corpus of general British English (and specifically, one which contains a range of texts that were published around the same time as our target corpus). The reference corpus would act as a benchmark for general English. Figure 8 shows how the output from CQPweb juxtaposes the frequency lists derived from the study corpus and the reference corpus. However, we might want to use a different 'comparator corpus' (Anderson, 2006) in order to draw out particular aspects of the target corpus. If we were interested in change over time, we might want to create a comparator corpus which contains similar articles about obesity, perhaps published ten years previously. Or if we wanted to focus on what is distinct about a particular newspaper, we could compare, say, a corpus of articles from the *Daily Mail* against a corpus of articles made up from a range of other newspapers.

The facility to create keywords is offered by most corpus tools, including WordSmith (Scott, 2020), AntConc, CQPweb, and Sketch Engine (and as listed in the Appendix, some of these tools come with accompanying pre-made reference corpora too). As with collocation, different tools offer a range of techniques for calculating keyness. Some techniques, like the log-likelihood



No.	Word	In whole "Obesity - News Corpus (version of May 2019)":		In corpus "50% sample of ukWaC":		+/-	Log likelihood
		Frequency (absolute)	Frequency (per mill)	Frequency (absolute)	Frequency (per mill)		
1	Obesity	64,673	1,793.82	10,004	8.88	+	391247.3
2	weight	69,353	1,923.63	83,089	73.72	+	277106.78
3	obese	38,780	1,075.63	2,967	2.63	+	248251.45
4	"	401,031	11,123.31	5,288,498	4,692.31	+	220298.88
5	fat	45,396	1,259.14	34,064	30.22	+	209062.97
6	she	118,162	3,277.43	740,356	656.89	+	179895.41
7	Sugar	38,264	1,061.32	27,741	24.61	+	177814.29
8	Diet	38,066	1,055.83	42,135	37.39	+	156185.66
9	Overweight	25,204	699.08	4,299	3.81	+	150898.34
10	diabetes	33,685	934.31	30,018	26.63	+	147852.96

Figure 8 Top ten keywords derived via CQPweb from comparing a corpus of UK news articles about obesity against ukWaC, a large reference corpus of British English internet texts

test, focus on the question of 'is there enough evidence that a word has a different relative frequency between two corpora?' Such techniques tend to favour reasonably high-frequency words (the higher the frequency, the more convincing the evidence), although they can also produce keywords where the frequency differences are not actually that large. Other techniques, like %DIFF, focus on the strength of difference as opposed to the amount of evidence (Gabrielatos and Marchi, 2011). So these techniques can favour lower-frequency words where the relative differences are larger. There is no 'best' way to calculate keyness, but it is useful to be aware of the types of keywords that will be produced. The %DIFF calculation is less likely than log-likelihood to elicit high-frequency grammatical words as keywords, for example.

Corpus analysis programs usually have default settings which will produce a number of keywords, although there are parameters that can be changed to produce different outputs. This search output usually ranks keywords in terms of their keyness score, with the ones at the top showing the most impressive differences in frequency. As with collocates, there is no ideal cut-off point and analysts must determine the optimal number of keywords to analyse, based on having enough to analyse while simultaneously avoiding the analysis becoming laborious to read.

Keywords can be useful for studying discourse because they help to reduce the large amount of language in a corpus down to a short list of more manageable words. Each keyword can function as an analytical signpost, revealing that a particular word is salient in a corpus and therefore worth focusing on. Because keywords are derived via statistical criteria, they are not subject to researcher bias, so they have the potential to direct analysts to words that they may not originally have considered to be particularly important. With that said, many keywords can be predicted in advance (a corpus of newspaper articles about obesity is likely to produce obvious keywords like *obese*, for example). The predictable keywords are likely to be ones that focus on important topics in the dataset, while the less predictable ones can sometimes be more revealing of stylistic differences. And it is the latter that might tell us how discourses or arguments are substantiated in more subtle ways.

In a keyword study (Baker, 2017) that compared personal adverts written by two sets of women (the first group seeking male partners, the second group seeking female partners), a keyword in the first group was *me*. A concordance analysis of this word indicated that it was most frequently used in constructions where the writer of the advertisement placed herself as the grammatical patient (i.e., the 'recipient') of their hypothetical romantic partner's actions (e.g., *I just want someone who can make me happy*). This appeared, then, to be a much more typical feature of women seeking men, compared to women seeking women. For this latter group, the keyword *we* indicated that they were more likely to picture themselves as engaged in shared activities with their hypothetical partner (e.g., *We can make dinner together*). The analysis thus reveals differences in how the two groups of women perceived their relationship to a new potential partner.

Another keyword study (Baker and Love, 2015) compared two sets of speeches made in the British Parliament relating to LGBT+ equality. The first set of speeches were made in 1998–2000 and related to reducing the age of consent to have sex for gay men to sixteen, to bring it into line with everyone else in the country. The second set of speeches were made in 2013 and related to the legalisation of same-sex marriage. Both corpora only contained speeches from politicians who were opposed to these changes to the law. A keyword for the 1998–2000 corpus was the pronoun *I*. This was found to be often used in instances where the speaker indicated their own belief or opinion (e.g., *I think*, *I believe*). A possible reason why the word *I* was key in the earlier corpus, then, was that speakers felt less concerned about being perceived by others as homophobic, so they were able to more explicitly 'own' their views by using the first-person pronoun. By contrast, during the time of the 2013 debate, societal attitudes had become more accepting of homosexuality (and there was more media scrutiny of politicians' opinions on the topic of LGBT+ equality), and so this might explain why the word *I* was used much less frequently in the later corpus of debates.

Even a predictable keyword can reveal points of interest. For example, a comparison of for-and-against arguments relating to foxhunting during a series of political debates found that speakers who wanted to ban foxhunting

https://doi.org/10.1017/9781009168144 Published online by Cambridge University Press

used the keyword *barbaric* (Baker, 2006). This was related to a moral argument where hunting was described as *barbaric*, along with similar words like *cruel*, *obscene*, and *bloodthirsty*. Such word choice is not all that surprising in anti-hunting arguments. What is interesting, however, is that a concordance analysis showed *barbaric* to refer to the practice of hunting rather than the people engaged in it. This is a subtle distinction and indicates a desire to appear not to personalise the debate, despite the fact that hunting is a practice carried out by individuals.

A good starting point for a keyness analysis is to try to group similar keywords into categories, then perhaps to select a few keywords for a more detailed analysis, via their concordance lines and collocates. This oscillation between different corpus linguistic tools is discussed further in Section 5. Two related questions to keep in mind are: What does this keyword achieve in this corpus? And why does that make this keyword particularly frequent in this corpus? It is not always necessary to analyse every word in a keyword list, particularly when this will result in repetition or reveal little of interest. As a matter of fact, simply working through a list of keywords one by one is likely to produce a disconnected analysis that will be tedious to read. Instead, analyses are likely to be most engaging if they focus on a set of connected keywords which contribute something original to a coherent and convincing narrative.

Depending on the cut-off points used, a keyword list can result in a very large number of candidate words for analysis, which will be difficult to do justice to. In such cases, selections should aim to avoid accusations of ‘cherry-picking’ as much as possible. For example, grouping keywords into themes, topics, or functional categories, then picking one or two of the more frequent keywords for more detailed analysis is likely to provide broader coverage and avoid analytical repetition. We would recommend that keywords which tell us something of unexpected or non-obvious meaning should take precedence over those which simply confirm what is already known. The analysis should ideally feature illustrative examples from the corpus texts as opposed to simply reciting frequencies and keyness scores. And it should be borne in mind that a *table* of keywords does not in itself constitute a keyword *analysis*. The table of keywords is the beginning of the analysis, or more accurately, the part prior to the analysis. The actual analysis is what happens when we try to work out why keywords appear, which involves consideration of context both within the corpus and outside it. For example, when Baker and Love (2015) studied the keyword *I* in the corpus of debates on LGBT+ equality, they considered relevant media commentary as well as the results of public attitude surveys towards homosexuality, and this background information helped them explain why the pronoun was key in the debates to equalise the age of consent.

https://doi.org/10.1017/9781009168144 Published online by Cambridge University Press

The keywords approach can also be extended to cover key multi-word units (referred to earlier as *n-grams*), key grammatical categories, or key semantic categories. For that purpose, different programs offer different facilities. Sketch Engine simultaneously produces a list of key multi-word units along with a list of keywords. Key multi-word units can also be obtained with WordSmith Tools, although the procedure is a little more complicated, requiring users to first make an index and then use this to derive a wordlist of clusters of a particular length. The online tool Wmatrix (Rayson, 2008) automatically assigns grammatical and semantic tags to each word in a corpus, and keyness calculations can then be performed on the frequencies of different tags as opposed to frequencies of different words. For example, the comparison of for and against arguments relating to foxhunting mentioned earlier found that speakers who wanted to keep foxhunting used more words that had been tagged with a semantic code for *Ethics*. This included words like *moral*, *ethical*, *rights*, and *humane* (Baker, 2006). Concordance analyses revealed that these words were used more by pro-hunting speakers because they engaged in more rhetorical work to question definitions of morality (e.g., *there are moral gradations here and no moral absolutes*).

Another important aspect to consider when conducting a keyword analysis is that it focuses on *differences* between two corpora. Differences can be interesting to analyse, but we ought to bear in mind that they may not be the most important aspects of a comparison; in fact, it is often the *similarities* between two corpora that are particularly worthy of discussion. There are a number of ways to investigate similarity: one way is to compare two corpora against a third reference corpus. Let's say, for example, that we are interested in looking at articles in the *Daily Mail* and *The Guardian* relating to obesity. A direct comparison would reveal keywords that indicate differences. However, comparing each corpus separately to a corpus of general British English would result in two keyword lists that we can then compare against each other. This side-by-side comparison might contain some words unique to one dataset, and some words shared between the two, thus helping us to spot both differences and similarities. Another approach could be to use a technique to identify lockwords (i.e., words where the relative frequencies are very similar), although currently only the web-based tool CQPweb allows one to do this. Taylor (2013) provides a helpful discussion of the importance of considering similarity alongside difference when carrying out CADS research.

Another potential restriction of the keywords approach is that it only allows two sets of corpora to be compared with each other. However, there are techniques which can be used to enable multiple comparisons. For example, Baker, Gabrielatos and McEnery (2013) examined a corpus of newspaper articles relating to Islam. The corpus contained articles from several newspapers, so in order to find

https://doi.org/10.1017/9781009168144 Published online by Cambridge University Press

out what was distinctive about the coverage within each newspaper, a series of keyword comparisons was carried out. In each instance, the articles from one of the newspapers studied – the *Daily Mail*, say – was compared to all the other articles from all the other newspapers in the corpus (*Guardian*, *The Sun*, *Daily Express*, etc.). This was repeated until one had a list of keywords for each newspaper. This method, referred to as the 'remainder method', can also be used to identify change over time, for example by dividing a diachronic corpus into different time periods and comparing each period with all the other periods in the corpus.

Key takeaways and things to watch out for:

- Keywords can act as analytical signposts, helping researchers to focus on what is most lexically salient in a corpus.
- At least two corpora are needed to obtain keywords, and the choice of reference corpus can impact on the keywords that are found – a point worth reflecting on in the analysis.
- Keywords reveal differences between *two* corpora. A workaround to deal with *multiple* comparisons is using the 'remainder method'.
- Analysis of keywords should involve identifying their collocates and/or viewing their concordance lines. We should ask what the keyword achieves in the corpus and why it might be used so frequently. To obtain a fuller explanation, we might need to draw on additional information from outside the dataset.
- The keywords technique can be expanded to involve key *n*-grams, key grammatical categories, or key semantic categories.

5 CADS in Practice

In the previous section, we introduced the corpus linguistic tools that are typically applied in corpus-assisted discourse studies (CADS). So now we know what tools are in the box, and how they work. But that alone doesn't tell us which tool is best suited at which point in the research process, and how the different tools are best orchestrated within a research design. There is no straightforward answer to these questions, and certainly none that will guarantee success every time. Yet some general guidance we feel can be given, and we would like to do so in two steps. First, we will use a brief example from an ongoing project to demonstrate how the analytical tools can be made to usefully interact. Second, we will offer a model for conceptualising the research process in a way that is more flexible than a simple flow chart.

5.1 Which Tool When? A Mini Case Study

To show how the tools introduced in Section 4 can work together, let's throw a spotlight on an ongoing project investigating dissent in judgements rendered by the UK Supreme Court (UKSC). Cases at the UKSC are decided by panels of judges. Those who do not agree with the majority of the other judges may write a so-called dissenting judgement (also referred to as *a dissent*), explaining what particular point they disagree on, and on what grounds. So why would dissents be of interest to the discourse analyst? It is because they matter – legally, politically, and socially – even though they have no immediate impact on the Court's final decision. For one thing, what is a dissent today may become the majority view in the future; for another, dissents open up a window to the discussion that the Court engaged in prior to passing judgement. Thus, a dissenting opinion isn't merely an act of disagreement; it is a text whose form and function are bound up with the legal framework the genre is embedded in. And far beyond the legal sphere, dissents are highly significant because they represent the very antithesis of violent conflict resolution, embodying instead a culture of balanced, reasoned debate. In fact, the genre epitomises a central democratic imperative: the twin principles of, on the one hand, giving divergent views a voice, while on the other allowing the majority view to prevail. To understand how this balance is struck, and how arguments are made, we need to unpack the role that language plays in 'doing' dissent. A small part of the analysis will be presented next.

As introduced in Section 4.1, the corpus consists of all the judgements from 2009 to 2018 which contain at least one dissenting opinion. It was split into two subcorpora: the majority judgements (MAJ) on the one hand, and the dissenting judgements (DISS) on the other. For a variety of reasons, we used Sketch Engine to carry out our analysis: it allows us to both upload our own corpora *and* make use of the corpora that it provides for contrastive analysis; it allows us to use the four corpus-assisted techniques explored in Section 4, but it also has other functionalities (such as the Word Sketch Difference tool) which may potentially come in useful throughout the analysis; and finally, it has a support function where you can request help from a member of their team – something that should not be taken for granted but is an incredibly useful option to have available.

Given our research question, it seemed reasonable to begin by exploring lexemes that could be expected to play a major part in expressing dissent. There is quite a range of lexical options, of course, some of which could be identified with the help of a thesaurus, while others might be opaque in the sense of not being easily identifiable on the linguistic surface. Yet for the purposes of our present demo, let's concentrate on one lexeme that is closely associated with disagreement, namely: *disagree*. Rather predictably, it is more frequent in the DISS than in the MAJ corpus: 92.57 occurrences per million tokens versus 64.05, respectively. This is the kind of underwhelming result which many researchers would probably choose not to mention, and those who do would have to deal with reviewers asking 'so what?' We do mention it here because distinguishing between findings that are worth reporting and those that are not can be a real challenge that we want to draw attention to. As it happens, in this particular instance, the result is not only unexciting, but it is also not terribly relevant. Because if we search for DISAGREE as a lemma,[5] the resulting concordance will include a fair number of instances that have nothing at all do with how judges frame dissent (e.g., *the two witnesses disagreed*). So how can we find what we're really after?

There are essentially two approach routes. One is via our knowledge of the conventions governing the genre in question, and that, in turn, can be gained by close reading of a fair number of texts from the corpus, as well as texts *about* the genre concerned, such as legal textbooks and treatises by academic lawyers. Reading and background research will have alerted us to the fact that British judges write judgements in their own name rather than as anonymous representatives of the Court (as is the case in Austria, for example). The first-person singular pronoun *I* is therefore a constant presence. If we are interested in how judges perform disagreement, then it is specifically *I disagree* that we need to look for (plus any synonyms which we may have decided to investigate alongside DISAGREE, including *cannot agree*). For only in the first person does *disagree* work as a performative speech act.

The second approach route – leading to the same conclusion – is to look at the collocations of DISAGREE. Here we used the MI statistic to calculate collocation, as we were interested in particularly strong collocates, likely to be distinctive of our specialised corpus. We defined a collocate as appearing three words either to the left or right of the node, and stipulated that it should appear at least five times in the corpus. There is nothing canonical about these choices, but experience suggested that they would be reasonable for the project at hand. In practice, one plays around with different settings, compares the output, and chooses the setting that appears to provide the most interesting perspective on the data.

[5] *Lemma* is defined as the 'base form of a word together with its inflected forms' (Collins, 2019: 197). Typically, lemmata are written in small caps.

That said, the output gleaned from other settings should not be dismissed offhand, and you may wish to return to alternative approaches later on.

And so, to return to our mini project, the collocations of DISAGREE are where things get rather more interesting. We can see that the gap between the two subcorpora, already noted with the presence of DISAGREE, begins to widen further. *I disagree* is more than twice as frequent in DISS than in MAJ (11.69 per million tokens vs. 5.74/m.); that trend continues upwards if we allow for an intervening adverb. The strongest collocate of DISAGREE in both MAJ and DISS, with MI scores of over 12, is *respectfully*, and that indeed is the adverb typically preceding DISAGREE. The phrase *I respectfully disagree* is more than three times as frequent in DISS (16.18/m.) than it is in MAJ (5.3/m.). Similarly, *with great respect* occurs much more frequently in DISS (8.09/m.) than in MAJ (2.65/m.). To return to our research question, then, one of the characteristics of judges' framing of dissent is to buffer its impact with standardised politeness markers.

Given that *respectfully* has emerged as a word of interest, a natural next step is to use it as a search term in its own right, illustrating the technique that Duguid and Partington (2018: 46) aptly refer to as 'chain concordancing'.[6] As one might expect, the quantitative evidence mirrors that related to DISAGREE: *respectfully* is nearly three times as frequent in DISS (45.83/m.) than in MAJ (19/m.). Examining the concordance, we can also see that the range of verbs following *respectfully* and describing a face-threatening speech act is wider in DISS than in MAJ. In the latter, in addition to *disagree*, we find only *differ* and *doubt*, whereas in the former, we find *doubt*, *question*, *reject*, and *take a different view*. That said, perhaps the most surprising discovery is that the second strongest collocate in the +1 position (i.e., to the immediate right of the node word) is AGREE. Thus, using Pomerantz's (1984) terminology, we can see that judges not only preface a *dispreferred* response with a face-saving politeness marker but also a *preferred* one, namely: agreement.

At first glance, this seems rather strange. Why buffer a speech act that to all intents and purposes is not face-threatening anyway? It seems obvious that the use of *respectfully* is heavily ritualised rather than an ad hoc, creative choice; and we can easily confirm that impression by looking at a reference corpus. The British National Corpus 1994 (BNC1994) tells us that *respectfully agree* is associated exclusively with law reports; in other words, the usage is domain- and genre-specific. However, we are still left with the question of why the ritual should be observed in the first place – that is, why speakers declare their respect when they are in agreement and, on the face of it, no offence has been

[6] Chain concordancing involves 'noting an interesting item in one concordance leading to its being concordanced in turn and so on, often involving several rounds' (Duguid and Partington, 2018: 46).

https://doi.org/10.1017/9781009168144 Published online by Cambridge University Press

committed. Wright et al. (2022) provide an interesting clue, commenting on the use of *respectfully* by the barrister Lord Pannick: ‘the source of the potential face-threat here, which is mitigated by *respect**,[7] is not *what* Pannick is saying to the court, but more likely that he is saying it at all’ (Wright et al., 2022: 10, italics added). In other words, any evaluative comment is a transgression of sorts, and *respectfully* serves as a general sign of deference.

Now that we have established that judges express dissent politely, we will want to explore what other options they use, apart from signalling their respect for those they disagree with. As we are likely to be dealing with a range of lexical realisations – each of them perhaps rather rare, but united by a common pragmatic function – this is the point where qualitative concordance analysis comes into its own. Still, we need a lexical hook on which we can hang our search. Any – or, better still, all – of our previous searches could serve that purpose. We’ll illustrate the principle by looking at a couple of concordance lines from the search for *respect* in the DISS subcorpus:

(1) I am afraid that, with respect, I must disagree.

(2) I have the greatest possible respect for the views of my colleagues and for the reasons which Lord Phillips has set out so carefully in his judgment. I regret however that I am unable to agree with what he proposes.

In both examples, several politeness markers are deployed, crammed into a very confined space: the apologetic preface *I am afraid that*; the modal *must* and *unable to* (suggesting that the speaker had no choice); an almost hyperbolic superlative (*the greatest possible respect*); and an explicit expression of regret. Clearly, these devices are mutually reinforcing. We could take each one of them and examine their concordances, eventually leading to a fairly comprehensive catalogue of the politeness markers used in this genre.

The point of this vignette was to give readers a glimpse of CADS ‘in action’: of finding entry points into the data; distinguishing promising paths from blind alleys; and interpreting the computer-generated output by drawing on evidence from outside the corpus itself. Above all, the case study showed how ‘one thing leads to another’ – a mundane, though effective technique, tried and tested in many a CADS project. For example, an unexpectedly frequent collocate may be a prime candidate for a concordance analysis, just as an outlier in a concordance may prompt further examination of frequency and dispersion. Thus, rather than using one corpus linguistic tool and ‘have done with it’ for the rest of the analysis, all tools are kept at the analyst’s disposal at all times and are activated

[7] The asterisk is a ‘wildcard’ which stands for all the possible word endings that *respect* can take (including *respect, respectfully, respecting,* etc.).

https://doi.org/10.1017/9781009168144 Published online by Cambridge University Press

as and when appropriate. The process is flexible and recursive, and we will provide a model for it in the next section.

Our demonstration of the research process should not be read uncritically as advice on how to write up research. In keeping with the genre of the academic textbook, we documented every phase of the process, letting readers in on our step-by-step reasoning. Yet such a painstaking logbook approach is not something that we would recommend for drafting research papers. When it's for real, a careful balance needs to be struck between, on the one hand, ensuring that irrelevant detail does not obscure the essence of your research story, and on the other, ensuring that your peers are given enough relevant detail so that they can make an informed judgement about the intellectual merits of your project.

5.2 From Single Voices to Polyphony

Our account so far – like those by many others writing about corpus linguistics (CL) – has probably suggested that there is a natural sequence in which CL analytics are best used. We may have implied, if only by the order in which we presented the tools, that frequency is the obvious starting point, to be followed by concordances, collocations, and then keyword analysis. Readers might easily have concluded that the tools are to be used one after the other. There is nothing wrong with this as such, and indeed many projects have proceeded in a sequential fashion. Nonetheless it seems worth questioning whether this protocol isn't merely simple but in fact oversimplified. For what researchers actually do is often rather less orderly – not because they do not know any better, but because complex data may require a more flexible approach, and because interim findings may open up unexpected vistas and heuristic opportunities. Many of these will of course be waiting to be further enriched by in-depth qualitative discourse analysis. After all, no social scientist would want to miss out on a chance to discover something new even if it means departing from the route originally planned. Sometimes a little messiness is a price worth paying (a viewpoint we will return to in Section 7).

Yet how can we embrace 'a little messiness' without jettisoning the idea of systematic and transparent data analysis? How, in other words, can we model a more adaptive use of tools without implying that 'anything goes'? To answer these questions, we would like to propose a musical metaphor. Suppose we regarded each tool as an instrument. In a sequential model, each instrument has its turn, but only once, and they never play together. In a quartet, by contrast (or a trio, or indeed a full orchestra, depending on how many tools you choose to employ), different instruments have their turn at different moments of the piece being played (i.e., the research process). On some occasions, two or more may

https://doi.org/10.1017/9781009168144 Published online by Cambridge University Press

be playing at the same time, while on others, one instrument may have a solo part. Whether separately or together, however, there must be no question that they are playing the same piece. In visual terms, the metaphor would give us something like Figure 9, a musical score.

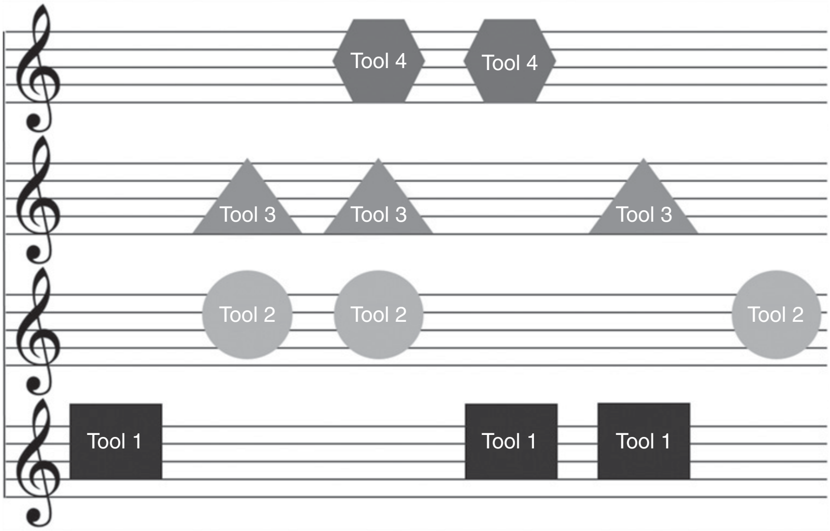


Figure 9 A schematic view of different CL tools being used at different stages of the research process

Thus, although Tool 1 (frequency) may be the starting point, we may want to return to it later (if Tool 3, collocation analysis, has produced a collocate whose frequency we are interested in); likewise, Tool 2 (concordances) may be used concurrently with all other tools (a process encouraged by Sketch Engine, for example, which allows one to switch directly from frequency lists and keywords to concordance lines of the items in question). In writing up our findings, we may still prefer to 'pretend' that all the tools were used at different stages – ostensibly for the sake of clarity and in the readers' interest, but perhaps also in response to pressure exerted by perceived epistemological norms (again an argument we will pick up on in Section 7). Whatever one's motive for presenting a sanitised version of the analytical process, we would alternatively propose telling research stories that are more fact than fiction.

The model presented in Figure 9 is in no way a radical departure. Much more mundanely, it simply acknowledges what is common practice anyway. And it might also make the modest, if substantive methodological point that CADS uses corpus tools flexibly, iteratively, and in a mutually reinforcing manner.

https://doi.org/10.1017/9781009168144 Published online by Cambridge University Press

6 Limitations and Potential Pitfalls

The previous sections should have made it abundantly clear why corpus-assisted discourse studies (CADS) is an attractive option. Our key selling point was that discourse studies (DS) can be enriched and made more rigorous by computer assistance. Yet we would be remiss if we tried to oversell that claim and did not acknowledge that the approach also has limitations. Equally, it seems important to alert readers to common pitfalls that they may encounter during their research journey.

6.1 Limitations

Like all other methods, CADS will not come into its own unless it is a good match for both the data and the specific research questions. For some types of projects, CADS will be an obvious choice; for others, it may be helpful but not essential; and for others still, it may not even be worth the effort. Methodological competence, then, consists not only in knowing when to apply a method and how, but also whether to apply it at all. The following remarks are intended to provide some guidance in that respect, and we can recommend Taylor and Marchi's (2018) *Corpus Approaches to Discourse: A Critical Review* for further discussion. So, while this Element so far has been about what CADS *can* do, we'll now have a look at what it can't do, or is less good at.

First, standard corpus linguistics (CL) programs are heavily biased towards lexis. We need words and phrases that we can search for, and we need a strong case for why these words and phrases play a key role in certain discourses. Because we always need a 'lexical hook', we will struggle if that does not exist or is impossible to identify with any certainty. This is likely to be the case whenever our focus is on broader discursive phenomena with multiple and unpredictable lexical realisations. Argumentative strategies or extended metaphors would be cases in point. For tackling these sorts of issues, we need to turn to traditional qualitative discourse analysis; in other words, the 'DS' part of 'CADS'.

Second, the focus on lexis means that the computer-assisted part of the analysis will tell us little about how meaning unfolds in longer stretches of text. Nor will it allow an in-depth exploration of how interactants negotiate meaning in conversation; or throw light on how the dynamics of social interaction evolve subtly over the course of a longer encounter.

Third, identifying phenomena that are absent from the corpus is another perennial problem, because corpus linguistic techniques, by their very nature, are primed to look for what is *there*. To counteract this, comparisons between corpora are essential. Duguid and Partington (2018: 56) show how such 'contrastive corpus techniques' can help show *what* is absent from a given corpus.

They do emphasise, however, that it remains the researcher's job, drawing upon their world knowledge, to interpret *why* something is absent.

Fourth, standard CL programs are biased towards verbal (i.e., textual) data. Removing language from its original context risks losing non-verbal information, such as photographs, that would otherwise be important to interpret the text as the author intended. Some impressive progress has been made in applying CL to multimodal data, including the analysis of images, sound, and gesture (Knight, 2011; Adolphs and Carter, 2013; Chen, Adolphs and Knight, 2020). In fact, Bednarek and Caple (2014: 151) propose a new method known as corpus-assisted multimodal discourse analysis (CAMDA), which combines a corpus-assisted analysis with the more fine-grained analysis of other semiotic modes (see also Bednarek and Caple, 2017; Caple, 2018; Caple, Huan and Bednarek, 2020). What is at the moment a highly specialised undertaking may in future become widespread within the CADS community at large.

Fifth, the available metadata may be insufficient to gauge the relevance of contextual factors, such as the age, gender, ethnicity, occupation, or relative social status of individual speakers or authors. If we have enough information to build corpora along such factors, we can frame our research questions accordingly (and ask, for example, how usage varies systematically with age or social class; see Baker and Brookes (2022) for further discussion). More metadata means more potential avenues of research.

Sixth, a word of caution is in order regarding CL's potential for reducing bias. There is the general proviso that no research is ever completely objective (Baker and McEnery, 2015: 9). Baker (2015: 286), for example, highlights how many decisions – ultimately subjective, however well-reasoned they may be – go into building research designs:

> As with interpretation and explanation, humans must make major methodological decisions at every stage: how to build a corpus; whether to annotate the corpus and, if so, with what annotation scheme; which software to use for analysis; which analytical procedures to implement and in which order; and which cut-off points for frequency or statistical saliency to apply.

In light of our comments throughout this section, one may wonder whether these six points should be referred to as *limitations* at all. We used the word in the section title, because that is what such reflective sections are commonly called. Yet perhaps that practice ought to be reconsidered. Is the wheel 'limited' for not helping us to fly? Is one of the 'limitations' of a fork that it is no good for eating soup? In that spirit, CADS should be judged against what it was designed to do in the first place. If a method appears to be limiting in some way, this need not mean that it is inherently flawed, but that it simply does not fit the research questions. Improve that fit, and the method will reach its full potential.

6.2 Potential Pitfalls

Building on our remarks on the limitations of CADS, we would now like to move on to the pitfalls that one may run into. We will be drawing on our own experience in a double capacity: partly as researchers who, like everyone else, have over the years had their fair share of mishaps and disappointments; and partly as reviewers who have seen quite a few submissions fail, not because the authors were lacking in technical competence but because they seemed to be unaware of the many other challenges posed by the CADS approach. Here, then, are the most important ones. Some arise in the corpus building phase, some in the analysis phase, and some in the writing-up phase. We will deal with each phase in turn.

During corpus building, there is the risk of data 'guzzling'. The comparative ease with which computerised data can be collected can encourage an indiscriminate 'all you can eat' approach to corpus building. Yet the available technology does not absolve one of the need to choose sources wisely; to reflect on how much and what kind of data is actually necessary to investigate a particular research question; and to carefully consider mundane but crucial questions around ethics and copyright. Texts should be collected because you need them, not because it is easy to do so.

During the analysis, further challenges await. The software generates wordlists, frequencies, concordances, and keyword statistics: reams and reams of evidence. But evidence of what? What does it all *mean*? The heart of a CADS inquiry is not an MI score or a logDice value; it is a plausible, well-argued connection between patterns of language use and the social or institutional context. In theses and journal submissions, it is not uncommon for the mechanics of the CL part to be executed impeccably, only for the paper to fall short at the stage of interpretation. It is at that point that you will inevitably come up against questions of ontology and epistemology (in very simplified terms, the nature of phenomena and how you can find out about them). What is the nature of the evidence that the software lays before you? What kinds of insight and knowledge does the method facilitate (and what kinds does it thwart)? And if you are part of a team, are you confident that all of the researchers involved in a particular project share the same ontological and epistemological assumptions (Ancarno, 2018; Jaworska and Kinloch, 2018)? These questions can provide valuable guidance, but only if they are addressed proactively and persistently.

https://doi.org/10.1017/9781009168144 Published online by Cambridge University Press

When struggling to make sense of one's data, it can be rather tempting to brush aside inconvenient or intractable evidence: for example, frequencies that are annoyingly at odds with what one was hoping to find; concordance lines that are difficult to interpret (Gillings and Mautner, forthcoming); and common collocates that simply do not seem to make sense. Such disappointing inconsistencies and contradictions cannot be argued out of existence, just as meaning cannot be plucked out of thin air through a categorically worded claim ('this shows that . . . :'; 'here we can see that . . . '). At this stage, therefore, it is crucial that interpretation proceeds cautiously and remains firmly grounded in the data. If at all possible, discuss your reading of the evidence with colleagues to see if it passes the litmus test of intersubjective validity.

More generally, there is always a danger that the very involvement of computers can lull us into a false sense of security, making us rely too much on them and be overconfident about our findings. There is thus the risk of an unintended irony. Methods originally introduced to alleviate bias may instead turn out to covertly aggravate it. Just as airbags can encourage speeding and thus, paradoxically, make driving more dangerous, corpus-assisted methods can tempt us to analyse our discourse data more carelessly than we would if we only had traditional, manual methods at our disposal. The most appropriate response to such doubts is not to reject corpus methods (or indeed airbags) but to make sure that their use is not regarded as carte blanche for reckless behaviour, which in scholarly terms would translate into flaws such as hasty overgeneralisation and unduly selective attention to some results but not others. Yet these flaws do not lie in the method, but in how it is applied.

In this context, it is also worth remembering that, to be worthy of its name, a CADS research design should include both corpus assistance and an orientation towards discourse. There seems to be a worrying trend to privilege the former over the latter. As a team from Northern Arizona University has recently found, corpus linguistic journal papers from 2019 dedicate more words to statistical reporting and less to linguistic description than those published in 2009 – a trend which they interpret as 'symptomatic of increasing distance from the language that is the object of study' (Larsson, Egbert and Biber, 2022: 137). We agree with Biber (2021), who reminds readers of the *Linguistics with a Corpus* blog that 'we have a lot to gain from staying close to the text'.

This is also true of the writing-up stage, in which a careful balance must be struck between reporting too little and reporting too much. Sweeping statements that leap directly from micro CL evidence to a macro discourse perspective are unconvincing. Yet so are overly detailed accounts that document every single step of the journey from raw data to interpretation, but lose sight of the original research question. In steering a middle course between these two extremes, it helps to heed the old advice, ‘write as a reader’. Put yourself in their shoes and then decide how much detail they are likely to need to assess the quality of your research, and how much they can take in without losing interest.

A key ingredient in maintaining that interest is to engage critically with your raw data as well as with the various outputs that the software generates. Being ‘critical’ means looking beyond what appears to be immediately obvious; acknowledging – always – that alternative readings might be possible; and stating clearly from which viewpoint one interprets the evidence. Critical distance is particularly important when the discourse under investigation is associated with, and shaped by, a political and social elite.

Overall, it is important to remember that basic principles of good research practice still apply – honed as they were a long time before computers were even invented: critical distance towards materials, methods, and findings; self-reflective critique; suitable caution in making far-reaching claims on the basis of limited evidence; and indeed a reader-centred style of writing that spells out arguments clearly. Above all, we must disabuse ourselves of the notion that any output, whether generated manually or by a machine, ever speaks for itself. The data does not speak; only researchers do.

7 Reflections on the Research Journey

We have now completed our *tour d’horizon* of corpus-assisted discourse studies (CADS): time to take stock of what we have learned and where we might go from here. As many of our sections have their own ‘Key takeaways and things to watch out for’ in a box at the end, it seems unnecessarily redundant to provide an additional summary here. Instead, we would like to take a step back, critically reflect on our own argument, and add a few thoughts that take us beyond the ‘tools and tips’ narrative that methods guides inevitably entail.

To begin, here is the traditional waiver, stated in the Introduction but worth repeating. Of course, this Element is incomplete. Our account of what CADS is and does had to be highly selective, not only for obvious reasons of space but also because a more elaborate treatise would have defeated the Element’s purpose as a compact and accessible invitation to engage with the methods proposed. Yet, while incomplete, we hope that the text will inspire researchers to integrate the approach into their own conceptual and methodological frameworks. And perhaps that inspiration can flourish precisely *because* the account is incomplete. For in a research landscape, it is often the gaps, nooks, and crannies where innovation springs up most readily.

https://doi.org/10.1017/9781009168144 Published online by Cambridge University Press

7.1 Dialogue Within and Across Disciplines

The specific strength of CADS no doubt derives from its mixed pedigree. It is part corpus linguistics and part discourse analysis, and that alone gives it a competitive edge over 'pure-bred' types of research design. Ideally, the empirical robustness of the corpus-assisted strand is combined with the insightful depth of the discourse analytical strand. After all, approaching a problem from two sides increases one's chances of solving it.

As we pointed out at the start of Section 2, the programmatic mixing that is wired into CADS makes it inherently triangulated (in theory, at least). Whether a specific project manages to successfully triangulate in practice is a different matter. As our experience as reviewers suggests, it is not uncommon for the 'CA' and 'DS' strands to remain quite separate, with only lip service being paid to the integration and synergies promised by the complete acronym. To return to the musical metaphor we introduced in Section 5, a key challenge in CADS consists in making sure that in the end we actually hear a complete piece of music. And that won't happen if the various instruments in the orchestra, however expertly played, just do their own thing without a conductor unifying them and guiding the interpretation.

That challenge becomes even tougher if CADS itself is only part of a broader repertoire of methods originating from disciplines other than linguistics. In that case, the mixed pedigree may involve entirely different species, which most likely use different terminologies and rely on different ontological and epistemological assumptions: about something as basic as the nature and role of language, for instance, or about how granular the analysis ought to be. (For example, Brookes and McEnery (2019) as well as Gillings and Hardie (2022) discuss this with regard to topic modelling.) At worst, such fundamental differences result in a haphazard collection of separate tools. At best, the differences lead to creative tension and an innovative, hybrid, and well-balanced ensemble of methods – precisely the flourishing in 'nooks and crannies' mentioned earlier.

Sadly and ironically, neither the utopian nor the dystopian vision of interdisciplinary cooperation is all that relevant in practical terms. For in reality, disciplinary divides remain strong, and often prohibitively so. And while there are notable examples of linguists collaborating with researchers from a range of other fields (Wodak, Kwon and Clarke, 2011; Pollach, 2012; Friginal, Mathews and Roberts, 2019; Collins et al., 2020; Dayrell, Ram-Prasad and Griffith-Dickson, 2020), such dialogue is still the exception rather than the rule; a matter of individual initiative rather than fully institutionalised cooperation. In Ancarno's (2018: 143) discussion of interdisciplinarity in

https://doi.org/10.1017/9781009168144 Published online by Cambridge University Press

CADS, it is suggested that ontological and epistemological beliefs may stand in the way of successful collaboration, as may the implications for the researcher's professional identity (either real or perceived), or lack of interest linked to so-called interdisciplinary fatigue. In the long run, a realistic solution could lie less in wishing disciplines away and more in backgrounding them. We are thus inclined to agree with Donaldson, Ward and Bradley (2010: 1534), who argue that '[a]cademic research as we know it currently cannot escape the shadow of disciplines, but in practice it can move towards ways of working in which the disciplines are not the most important things at play'. In the end, disciplinary labels and identities ought to matter less than the commitment to unravel the mysteries of language – a phenomenon so complex, after all, that we must throw everything at it. Chafe's (1992: 96) passionate plea, made over thirty years ago, has lost none of its relevance and appeal.

> I continue to believe that one should not characterize linguists, or researchers of any kind, in terms of a single favorite tie to reality. […] I would like to see the day when we will all be more versatile in our methodologies, skilled at integrating all the techniques we will be able to discover for understanding this most basic, most fascinating, but also most elusive manifestation of the human mind.

In that spirit, CADS is well placed to make a useful contribution.

7.2 Current Developments and Future Directions

CADS is a lively and growing field. One recent area of development revolves around the ways in which keywords can be calculated. For example, Egbert and Biber (2019) have argued that many keywords are not widely dispersed across a corpus, so are not truly representative of the text that they are derived from. Instead, they have suggested a new way of identifying keywords, based on text dispersion rather than frequency. Gries (2021) extends this idea by adopting a two-dimensional approach which takes into account both frequency and dispersion. Additionally, there have been discussions about the most appropriate measure for calculating keyness. Tests of statistical significance, like the log-likelihood measure, have faced criticism for over a decade; for example, Gabrielatos and Marchi (2011) have argued that log-likelihood prioritises keywords that have high *absolute* frequency, but demonstrate quite small differences in *relative* frequency. As an alternative metric, they have recommended effect-size measures like %DIFF (see also Gabrielatos, 2018), while Hardie (2014) suggests an effect-size statistic called *Log Ratio*. More recently, Jeaco (2020) compared different measures for calculating keyness and concluded that the use of log-likelihood is fit for purpose, particularly if coupled with Bayes factors. Many corpus tools now

https://doi.org/10.1017/9781009168144 Published online by Cambridge University Press

offer a wide range of keyness measures so that the researcher can determine which is the most appropriate for their specific project. To give but one example, #LancsBox offers five techniques to choose from (i.e., simple maths, %DIFF, log-likelihood, Log Ratio, and Cohen's D).

Another strand of research has been to identify automatic ways of categorising keywords. For example, Clarke, McEnery and Brookes (2021) have used Multiple Correspondence Analysis to not only group similar keywords in a corpus but also to identify sub-registers or discourses based on texts that contain similar keywords. The advantages of the approach are not only that it can be used to show changes in discourse through time (Clarke, Brookes and McEnery, 2022) but also that it can be carried out on corpora that contain very short texts, such as tweets. The proliferation of new techniques will hopefully result in more insightful analyses, although they may also result in confusion, especially for beginners, along with the evident difficulties among researchers to reach a consensus about the most effective procedures. Additionally, such techniques need to be incorporated into corpus analysis programs; otherwise they are unlikely to be adopted widely.

There has also been a move away from off-the-shelf corpus analysis programs (such as those listed in the Appendix), and towards the use of the programming language R (Gries, 2009; Winter, 2019). This approach frees the researcher from the constraints of existing tools, enabling more complex and sophisticated forms of analyses to be carried out. The difficulty here, though, is that it requires CADS researchers to take the time to learn and develop statistical and programming skills, which many may not be able, or inclined, to do.

Another trend has involved the spread of CADS research to languages other than English. For example, Thornborrow, Ekstrom and Patrona (2021) examined the discursive representation of populism in newspapers from Greece, Sweden, France, and the United Kingdom. Investigating German news articles, Dykes and Peters (2020) examined argumentation patterns about drug-resistant pathogens. And in a Brazilian context, Rebechi (2019) analysed political speeches relating to the impeachment of President Rousseff in the country's Lower House of Congress. Other forms of CADS research have engaged with texts in English while considering contexts beyond the United Kingdom. For example, Liu (2022) examined stance-taking in the speeches of three former chief executives in Hong Kong, and Pei, Li and Cheng (2022) looked at the representation of hackers in corpora consisting of news articles from *China Daily* and the *New York Times*.

While it is clear that CADS is indeed expanding beyond the British and European academic context, we wish to highlight that many researchers

https://doi.org/10.1017/9781009168144 Published online by Cambridge University Press

around the world face systematic constraints that limit their ability to engage in CADS as freely as they may wish. One set of constraints is institutional and financial. It is undeniable that many colleagues are placed at a disadvantage because they do not have equal access to corpus analysis programs, corpora in English and other languages, research funding, academic literature, or opportunities to exchange ideas at conferences and workshops. Recent advances in the open access movement, twinned with the continued updating of free corpus analysis programs, are helping matters – but more could certainly be done. After all, corpus-assisted techniques can be applied to any language or cultural context, and it is hoped that the field will broaden its remit further in the coming decades.

A second set of possible constraints is political. The more authoritarian a political system, the more difficult it is to engage in the kind of analysis and debate that pluralistic academic cultures take for granted. The required critical distance tends not to cause problems in the corpus-assisted strand of the CADS research process, but it often becomes an issue in the equally important discourse-analytic strand. In the latter, findings have to be contextualised and interpreted – during those stages of the process, in other words, which invariably bring the researcher up against questions of values, ideology, and power (whether or not their research carries an explicitly ‘critical’ label). As we have pointed out before, engaging in CADS is about ‘DS’ as much as about ‘CA’, and the need to address both with equal rigour and clarity is non-negotiable.

7.3 Navigating Mess and a Craft Attitude to Research

Acknowledging that CADS is an exercise in triangulation and works across disciplinary divides does not automatically translate into an infallible research protocol. The question ‘how does one actually do it?’ thus persists. Some researchers reject the very idea of a protocol because it strikes them as very rigid; others embrace protocols for the same reason. A third group, including the present authors, sees value in carefully planning research designs, while also acknowledging that there needs to be room for flexibility and intuition. To account for the inevitable messiness of discourse, we should, by all means, play with the data, but we should play with a plan.

Throughout this Element, we have tried to address the ‘how to’ question, and at the same time embedded the answers in suitable theoretical frameworks. We offered a worked example, referred readers to published works in the CADS tradition, and introduced a metaphor, the musical score, for dealing with the flexible entrances, silences, and exits of different research instruments. Ultimately, however, none of this can prepare you for your encounter with

https://doi.org/10.1017/9781009168144 Published online by Cambridge University Press

real data, whether playful or structured: the sheer mass of text, meaning it is difficult to know where to start; the often futile search for patterns that one was hoping to find; and the disconcerting surprises that occur when the corpus linguistic evidence will not properly align with the discourse-analytic diagnosis, and vice versa. On the plus side, there may be equally unexpected but very welcome discoveries; patterns that emerge where one least expected them, and which manual analysis would likely not have spotted. So it is not all doom and gloom, but one should not expect CADS to provide instant enlightenment.

Naturally, as in all research endeavours, there is a temptation to see only what one wants to see – not as the result of a lapse in academic integrity and even less as a wilful act of deception, but simply because the human eye is trained to accommodate to objects by adjusting its lens. The researcher's analytical eye is no different. Can we guard against the resulting distortion? Up to a point, yes. Corpus-assisted discourse analyses and other triangulated research designs are intended to provide as balanced a view as possible. A self-reflexive attitude to the research process should ensure that evidence is neither over- nor under-interpreted, and that 'blind spots' and 'dusty corners' are suitably addressed (Marchi and Taylor, 2018). At any rate, the problem is not that research may give you a distorted view of reality, but that we often harbour the illusion that an objective view is even possible. As a matter of fact, '[a]ll research is performative, in the sense that it enacts the real' (Markussen, 2005: 329). Methods, then, are usefully conceptualised as 'methodological interventions' (Davies and Dwyer, 2007: 262). So perhaps paradoxically, we can deal with distortion most effectively by first acknowledging that it is inevitable and then putting appropriate checks and balances in place. Thus, the research process is best seen as a way of navigating mess rather than eliminating it.

Since Law's 2004 book, *After Method: Mess in Social Science Research*, mess has acquired a certain respectability. Not to the extent of being welcomed, exactly, but as something that is to be accepted and not brushed under the carpet. In the social sciences and with a view to mixed-methods research, authors now openly acknowledge that 'epistemological norms pressure researchers to present unified, singular views of reality' (Sanscartier, 2020: 48–9). In DS, we would argue, we are faced with a similar situation.

CADS projects potentially involve both kinds of mess that Sanscartier (2020) identifies: 'empirical mess' and 'design-related mess'. The former arises 'when findings do not hang together neatly, or fail to produce a coherent, single picture of reality' (p. 48), a typical instance being quantitative and qualitative findings that do not converge. Design-related mess occurs when researchers have to adapt to unexpected circumstances, following paths not quite as neat and linear

https://doi.org/10.1017/9781009168144 Published online by Cambridge University Press

as methods textbooks may suggest (p. 49). For CADS practitioners, both scenarios will sound frustratingly familiar.

Fortunately, the solution proposed by the mixed-methods literature will also resonate strongly with anyone involved in CADS (Johnson, 2011; Maxwell, 2013). To deal with mess, the argument goes, the researcher needs to become what Levi-Strauss (1962, 1996) called a *bricoleur*, ‘a methodological “handy-person” capable of adaptation and improvisation’ (Sanscartier, 2020: 50). Underpinning these qualities is a ‘craft attitude’, which Sanscartier (2020: 53), drawing on Daft (1983), defines as follows:

> a disposition (not a paradigm, method, or design type) towards the mixed methods research process that (a) is comfortable with uncertainty, (b) favors non-linearity and recursiveness in research design, and (c) treats research as an exercise in storytelling, about both the research object and our engagement with that object.

This definition echoes several themes which we have touched upon in this Element. We talked about the spurious sense of certainty that the use of computer-assisted tools may create; we highlighted the need to combine tools flexibly at different stages in the research design rather than deploying them in a fixed and linear sequence; and we discussed the value of framing one’s findings as part of a research *process* rather than as neatly organised *products*. Needless to say, the craft principle can be (ab)used to hide a multitude of sins: the random choice of methods instead of their principled selection; flawed reasoning passed off as insightful intuition; and mess not just thoughtfully accepted, but negligently created. Yet any principle can be taken to pointless extremes and applied dysfunctionally, within qualitative as well as quantitative research designs. So the potential for abuse should not in itself put us off dealing with mess proactively and purposefully.

Like the *bricoleur*, CADS researchers content themselves with creating knowledge that is partial, fluid, and heavily context-dependent. They accept that their research journey may sometimes turn out to be rambling and in the end even circular. Yet, as T. S. Eliot (2001 [1943]) reminds us in his poem ‘Little Gidding’, a quest is not futile simply because it turns out to come full circle in the end:

> We shall not cease from exploration
> And the end of all our exploring
> Will be to arrive where we started
> And know the place for the first time.

The journey itself can thus change our perspective of what we went out to investigate. Ultimately, the value of CADS, too, lies in its power to transform.

Appendix: Corpus Analysis Programs

All the corpus analysis programs outlined in Table A1 broadly offer the same tools: frequency, concordances, collocation, and keywords (plus selected other features such as *n*-grams). However, the user interface will differ, as may the names given to these tools and the parameters that one can set within them. The list is not complete, and we have selected those programs that are most commonly used for corpus-assisted discourse studies (CADS). Readers may wish to conduct their own searches to find additional programs online.

https://doi.org/10.1017/9781009168144 Published online by Cambridge University Press

https://doi.org/10.1017/9781009168144 Published online by Cambridge University Press

Program	Corpora provided/upload your own	Free/paid	Online/download
WordSmith Tools (Scott, 2020)	Upload your own	Paid	Download
CQPweb (Hardie, 2012)	Both	Free	Online
AntConc (Anthony, 2022)	Both	Free	Download
#LancsBox (Brezina, McEnery and Wattam, 2015)	Both	Free	Download
Sketch Engine (Kilgarriff et al., 2014)	Both	Paid	Online
English-Corpora.org	Corpora provided	Both	Online
Wmatrix (Rayson, 2008)	Both	Depends on affiliation	Online

Table A1 Comparison of corpus analysis programs

References

Adolphs, S. and Carter, R. (2013). *Spoken Corpus Linguistics: From Monomodal to Multimodal*. London: Routledge.

Ancarno, C. (2018). Interdisciplinary approaches in corpus linguistics and CADS. In C. Taylor and A. Marchi, eds., *Corpus Approaches to Discourse: A Critical Review*. London: Routledge, pp. 130–55.

Ancarno, C. (2020). Corpus-assisted discourse studies. In A. De Fina and A. Georgakopoulou, eds., *The Cambridge Handbook of Discourse Studies*. Cambridge: Cambridge University Press, pp. 165–85.

Anderson, W. J. (2006). *The Phraseology of Administrative French: A Corpus-Based Study*. Amsterdam: Rodopi.

Anthony, L. (2022). *AntConc* (Version 4.1.1) [Computer Software]. Tokyo: Waseda University. www.laurenceanthony.net/software.

Appleton, S. (2021). East, West, and Westminster: A corpus-based study of UK foreign policy statements regarding the unification of Germany, October 1989 to November 1990. *Journal of Corpora and Discourse Studies*, 4: 39–62.

Archer, D. and Gillings, M. (2020). Depictions of deception: A corpus-based analysis of five Shakespearean characters. *Language and Literature*, 29(3): 246–74.

BAAL [British Association for Applied Linguistics] (2021). Recommendations on Good Practice in Applied Linguistics. www.baal.org.uk/wp-content/uploads/2021/03/BAAL-Good-Practice-Guidelines-2021.pdf.

Baker, P. (2006). *Using Corpora in Discourse Analysis*. London: Continuum.

Baker, P. (2014). *Using Corpora to Analyse Gender*. London: Bloomsbury.

Baker, P. (2015). Does Britain need any more foreign doctors? Inter-analyst consistency and corpus-assisted (critical) discourse analysis. In N. Groom, M. Charles and S. John, eds., *Corpora, Grammar and Discourse: In Honour of Susan Hunston*. Amsterdam: John Benjamins, pp. 283–300.

Baker, P. (2016). The shapes of collocation. *International Journal of Corpus Linguistics*, 21(2): 139–64.

Baker, P. (2017). Sexuality. In E. Friginal, ed., *Studies in Corpus-Based Sociolinguistics*. London: Routledge, pp. 159–77.

Baker, P. and Baker, H. (2019). Conceptualising masculinity and femininity in the British press. In C. Carter, L. Steiner and S. Allan, eds., *Journalism, Gender and Power*. London: Routledge, pp. 363–82.

Baker, P. and Brookes, G. (2022). *Analysing Language, Sex and Age in a Corpus of Patient Feedback: A Comparison of Approaches*. Cambridge: Cambridge University Press.

Baker, P., Brookes, G. and Evans, C. (2019). *The Language of Patient Feedback: A Corpus Linguistic Study of Online Health Communication*. London: Routledge.

Baker, P. and Egbert, J. (eds.). (2016). *Triangulating Methodological Approaches in Corpus-Linguistic Research*. London: Routledge.

Baker, P. and Ellece, S. (2011). *Key Terms in Discourse Analysis*. London: Continuum.

Baker, P., Gabrielatos, C., KhosraviNik, M., Krzyżanowski, M., McEnery, T. and Wodak, R. (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.

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.

Baker, P. and Love, R. (2015). The hate that dare not speak its name? *Journal of Language, Aggression and Conflict*, 3(1): 57–86.

Baker, P. and McEnery, T. (eds.). (2015). *Corpora and Discourse Studies: Integrating Discourse and Corpora*. Basingstoke: Palgrave Macmillan.

Baker, P., Vessey, R. and McEnery, T. (2021). *The Language of Violent Jihad*. Cambridge: Cambridge University Press.

Bednarek, M. and Caple, H. (2014). Why do news values matter? Towards a new methodological framework for analysing news discourse in critical discourse analysis and beyond. *Discourse and Society*, 25(2): 135–58.

Bednarek, M. and Caple, H. (2017). *The Discourse of News Values: How News Organizations Create Newsworthiness*. New York: Oxford University Press.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, 8(4): 243–57.

Biber, D. (2015). Corpus-based and corpus-driven analyses of language variation and use. In B. Heine and H. Narrog, eds., *The Oxford Handbook of Linguistic Analysis*, 2nd ed. Oxford: Oxford University Press, pp. 159–91.

Biber, D. (2021). *Corpus Linguistics Is for Text Lovers*. Linguistics with a Corpus [Companion blog to Egbert, Larsson and Biber, 2020], 22 December. https://linguisticswithacorpus.wordpress.com/2021/12/22/corpus-linguistics-is-for-text-lovers%EF%BF%BC/.

Biber, D., Reppen, R., Schnur, E. and Ghanem, R. (2016). On the (non)utility of Juilland's D to measure lexical dispersion in large corpora. *International Journal of Corpus Linguistics*, 21(4): 439–64.

https://doi.org/10.1017/9781009168144 Published online by Cambridge University Press

Brezina, V. (2018). *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge: Cambridge University Press.

Brezina, V., Hawtin, A. and McEnery, T. (2021). The Written British National Corpus 2014: Design and comparability. *Text & Talk*, 41(5–6): 595–615.

Brezina, V., McEnery, T. and Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, 20(2): 139–73.

Brezina, V. and Meyerhoff, M. (2014). Significant or random? A critical review of sociolinguistic generalisations based on large corpora. *International Journal of Corpus Linguistics*, 19(1): 1–28.

Brookes, G. and Baker, P. (2021). *Obesity in the News: Language and Representation in the Press*. Cambridge: Cambridge University Press.

Brookes, G. and McEnery, T. (2019). The utility of topic modelling for discourse studies: A critical evaluation. *Discourse Studies*, 21(1): 3–21.

Caldas-Coulthard, C. R. (1993). From discourse analysis to critical discourse analysis: The differential re-presentation of women and men speaking in written news. In J. M. Sinclair, M. Hoey and G. Fox, eds., *Techniques of Description: Spoken and Written Discourse*. London: Routledge, pp. 196–208.

Caple, H. (2018). Analysing the multimodal text. In C. Taylor and A. Marchi, eds., *Corpus Approaches to Discourse: A Critical Review*. London: Routledge, pp. 85–109.

Caple, H., Huan, C. and Bednarek, M. (2020). *Multimodal News Analysis Across Cultures*. Cambridge: Cambridge University Press.

Chafe, W. (1992). The importance of corpus linguistics to understanding the nature of language. In J. Svartvik, ed., *Directions in Corpus Linguistics: Proceedings of the Nobel Symposium 82, Stockholm, 4–8 August 1991*. Berlin: Mouton de Gruyter, pp. 79–97.

Chen, Y., Adolphs, S. and Knight, D. (2020). Multimodal discourse analysis. In E. Friginal and J. Hardy, eds., *The Routledge Handbook of Corpus Approaches to Discourse Analysis*. London: Routledge, pp. 98–115.

Cheng, W. and Lam, P. W. Y. (2013). Western perceptions of Hong Kong ten years on: A corpus-driven critical discourse study. *Applied Linguistics*, 34(2): 173–90.

Clarke, I., Brookes, G. and McEnery, T. (2022). Keywords through time: Tracking changes in press discourses of Islam. *International Journal of Corpus Linguistics*, 27(4): 399–427.

Clarke, I., McEnery, T. and 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): 144–71.

https://doi.org/10.1017/9781009168144 Published online by Cambridge University Press

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

Collins, L. and Hardie, A. (2022). Making use of transcription data from qualitative research within a corpus-linguistic paradigm: Issues, experiences, and recommendations. *Corpora*, 17(1): 123–35.

Collins, L., Semino, E., Demjén, Z., Hardie, A., Moseley, P., Woods, A. and Alderson-Day, B. (2020). A linguistic approach to the psychosis continuum: (Dis)similarities and (dis)continuities in how clinical and non-clinical voice-hearers talk about their voices. *Cognitive Neuropsychiatry*, 25(6): 447–65.

Culpeper, J. and Gillings, M. (2018). Politeness variation in England: A north-south divide? In V. Brezina, R. Love and K. Aijmer, eds., *Sociolinguistic Studies of the Spoken BNC2014*. London: Routledge, pp. 33–59.

Culpeper, J., Hardie, A., Demmen, J., Hughes, J. and Timperley, M. (2021). Supporting the corpus-based study of Shakespeare's language: Enhancing a corpus of the First Folio. *ICAME Journal*, 45(1): 37–86.

Daft, R. L. (1983). Learning the craft of organizational research. *Academy of Management Review*, 8: 539–46.

Davies, M. (2008). *The Corpus of Contemporary American English (COCA)*. www.english-corpora.org/coca/.

Davies, G. and Dwyer, C. (2007). Qualitative methods: Are you enchanted or are you alienated? *Progress in Human Geography*, 31(2): 257–66.

Dayrell, C., Ram-Prasad, C. and Griffith-Dickson, G. (2020). Bringing corpus linguistics into religious studies: Self-representation amongst various immigrant communities with religious identity. *Journal of Corpora and Discourse Studies*, 3: 96–121.

Dayter, D. and Messerli, T. C. (2022). Persuasive language and features of formality on the r/ChangeMyView subreddit. *Internet Pragmatics*, 5(1): 165–95.

Denzin, N. K. (1970). *The Research Act in Sociology*. Chicago: Aldine.

Donaldson, A., Ward, N. and Bradley, S. (2010). Mess among disciplines: Interdisciplinarity in environmental research. *Environment and Planning A: Economy and Space*, 42(7): 1521–36.

Duguid, A. and Partington, A. (2018). Absence: You don't know what you're missing. Or do you? In C. Taylor and A. Marchi, eds., *Corpus Approaches to Discourse: A Critical Review*. London: Routledge, pp. 38–59.

Durrant, P. and Doherty, A. (2010). Are high frequency collocations psychologically real? Investigating the thesis of collocational priming. *Corpus Linguistics and Linguistic Theory*, 6(2): 125–55.

Dykes, N. and Peters, J. (2020). Reconstructing argumentation patterns in German newspaper articles on multidrug-resistant pathogens: A multi-measure keyword approach. *Journal of Corpora and Discourse Studies*, 3: 51–74.

Egbert, J. and Baker, P. (eds.). (2020). *Using Corpus Methods to Triangulate Linguistic Analysis*. London: Routledge.

Egbert, J. and Biber, D. (2019). Incorporating text dispersion into keyword analyses. *Corpora*, 14(1): 77–104.

Egbert, J., Larsson, T. and Biber, D. (2020). *Doing Linguistics with a Corpus: Methodological Considerations for the Everyday User*. Cambridge: Cambridge University Press.

Egbert, J. and Schnur, E. (2018). The text in corpus and discourse analysis. In C. Taylor and A. Marchi, eds., *Corpus Approaches to Discourse: A Critical Review*. London: Routledge, pp. 160–73.

Eliot, T. S. (2001 [1943]). *Four Quartets*. Eastbourne: Gardners Book, Harcourt.

Evison, J. (2010). What are the basics of analysing a corpus? In A. O'Keeffe and M. McCarthy, eds., *The Routledge Handbook of Corpus Linguistics*. London: Routledge, pp. 122–35.

Fairclough, N. (1992). *Discourse and Social Change*. Cambridge: Polity Press.

Firth, J. R. (1935). The technique of semantics. *Transactions of the Philological Society,* 34(1): 36–73.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930–1955. In Philological Society, ed., *Studies in Linguistic Analysis*. Oxford: Blackwell, pp. 1–32.

Fitzgerald, C. (2020). Penetrating historical discourse's truth matrix: A corpus analysis of oral history testimonies. *Journal of Corpora and Discourse Studies*, 3: 75–95.

Formato, F. (2019). *Gender, Discourse and Ideology in Italian*. Cham: Palgrave Macmillan.

Friginal, E. and Hardy, J. (eds.). (2021). *The Routledge Handbook of Corpus Approaches to Discourse Analysis*. London: Routledge.

Friginal, E., Mathews, E. and Roberts, J. (2019). *English in Global Aviation: Context, Research, and Pedagogy*. London: Bloomsbury.

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): 846–85.

Gabrielatos, C. (2007). Selecting query terms to build a specialised corpus from a restricted-access database. *ICAME Journal*, 31: 5–44.

Gabrielatos, C. (2018). Keyness analysis: Nature, metrics and techniques. In C. Taylor and A. Marchi, eds., *Corpus Approaches to Discourse: A Critical Review*. London: Routledge, pp. 225–58.

Gabrielatos, C. and Marchi, A. (2011). Keyness: Matching metrics to definitions. Corpus Linguistics in the South 1, University of Portsmouth, 5 November.

Gillings, M. (2021). A corpus-based investigation into verbal cues to deception and their sociolinguistic distribution. Unpublished PhD thesis, Lancaster University.

Gillings, M. (2022). How to use corpus linguistics in forensic linguistics. In A. O'Keeffe and M. McCarthy, eds., *The Routledge Handbook of Corpus Linguistics*, 2nd ed. London: Routledge, pp. 589–601.

Gillings, M. and Dayrell, C. (2023). Climate change in the UK press: Examining discourse fluctuation over time. *Applied Linguistics*. https://doi.org/10.1093/applin/amad007.

Gillings, M. and Hardie, A. (2022). The interpretation of topic models for scholarly analysis: An evaluation and critique of current practice. *Digital Scholarship in the Humanities*. https://doi.org/10.1093/llc/fqac075.

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

Gomide, A. R. (2020). *Corpus linguistics software: Understanding their usages and delivering two new tools*. PhD Thesis, Lancaster University. https://eprints.lancs.ac.uk/id/eprint/149537/.

Gries, S. T. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, 13(4): 403–37.

Gries, S. T. (2009). *Statistics for Linguistics with R: A Practical Introduction*. Berlin: Mouton de Gruyter.

Gries, S. T. (2015). The most under-used statistical method in corpus linguistics: Multi-level (and mixed-effects) models. *Corpora*, 10(1): 95–125.

Gries, S. T. (2021). A new approach to (key) keywords analysis: Using frequency, and now also dispersion. *Research in Corpus Linguistics*, 9(2): 1–133.

Gupta, K. (2015). *Representation of the British Suffrage Movement*. London: Bloomsbury.

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

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

Hardie, A. (2014). *Log Ratio: An Informal Introduction*. Post on the website of the ESRC Centre for Corpus Approaches to Social Science (CASS). http://cass.lancs.ac.uk/?p=1133.

Hardt-Mautner, G. (1995). 'Only connect': Critical discourse analysis and corpus linguistics. UCREL Technical Paper 6, Lancaster University. http://ucrel.lancs.ac.uk/papers/techpaper/vol6.pdf.

Harvey, R. (2020). Twitter reactions to the UN's #HeForShe campaign for gender equality: A corpus-based discourse analysis. *Journal of Corpora and Discourse Studies*, 3: 31–50.

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Hunston, S. (2022). How can a corpus be used to explore patterns? In A. O'Keeffe and M. McCarthy, eds., *The Routledge Handbook of Corpus Linguistics*, 2nd ed. London: Routledge, pp. 140–54.

Hunt, S. (2015). Gender and agency in the Harry Potter series. In P. Baker and T. McEnery, eds., *Corpora and Discourse Studies: Integrating Discourse and Corpora*. Basingstoke: Palgrave Macmillan, pp. 266–84.

Hunt, D. and Brookes, G. (2020). *Corpus, Discourse and Mental Health*. London: Bloomsbury.

Jaworska, S. (2018). Change but no climate change: Discourses of climate change in corporate social responsibility reporting in the oil industry. *International Journal of Business Communication*, 55(2): 194–219.

Jaworska, S. and Kinloch, K. (2018). Using multiple data sets. In C. Taylor and A. Marchi, eds., *Corpus Approaches to Discourse: A Critical Review*. London: Routledge, pp. 110–29.

Jaworski, A. and Coupland, N. (ed.). (2014). *The Discourse Reader*. London: Routledge.

Jeaco, S. (2020). Key words when text forms the unit of study: Sizing up the effects of different measures. *International Journal of Corpus Linguistics*, 25 (2): 125–54.

Johnson, R. B. (2011). Do we need paradigms? A mixed methods perspective. *Mid-Western Educational Researcher*, 11(2): 156–73.

Joulain-Jay, A. (2017). *Corpus linguistics for history: The methodology of investigating place-name discourses*. PhD Thesis, Lancaster University. https://eprints.lancs.ac.uk/id/eprint/88671/.

Juilland, A. G., Brodin, D. R. and Davidovitch, C. (1970). *Frequency Dictionary of French Words*. The Hague: Mouton de Gruyter.

Kilgarriff, A., Baisa, V., Bušta, J. Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, 1: 7–36.

Knight, D. (2011). *Multimodality and Active Listenership: A Corpus Approach*. London: Bloomsbury.

Knight, D. and Adolphs, S. (2022). Building a spoken corpus: What are the basics? In A. O'Keeffe and M. McCarthy, eds., *The Routledge Handbook of Corpus Linguistics*, 2nd ed. London: Routledge, pp. 21–34.

Knight, D., Morris, S., Arman, L., Needs, J. and Rees, M. (2021). *Building a National Corpus: A Welsh Language Case Study*. London: Palgrave Macmillan.

Koester, A. (2022). Building small specialised corpora. In A. O'Keeffe and M. McCarthy, eds., *The Routledge Handbook of Corpus Linguistics*, 2nd ed. London: Routledge, pp. 48–61.

Koller, V. (2007). 'The world's local bank': Glocalisation as a strategy in corporate branding discourse. *Social Semiotics*, 17(1): 111–31.

Kopf, S. (2019). Content policies in social media critical discourse studies: The invisible hand of social media providers? *Critical Approaches to Discourse Analysis Across Disciplines*, 11(1): 1–19.

Krendel, A., McGlashan, M. and Koller, V. (2022). The representation of gendered social actors across five manosphere communities on Reddit. *Corpora*, 17(2): 291–321.

Larsson, T., Egbert, J. and Biber, D. (2022). On the status of statistical reporting versus linguistic description in corpus linguistics: A ten-year perspective. *Corpora*, 17(1): 137–57.

Law, J. (2004). *After Method: Mess in Social Science Research*. London: Routledge.

Leech, G. and Fallon, R. (1992). Computer corpora: What do they tell us about culture? *ICAME Journal*, 16: 29–50.

Levi-Strauss, C. (1962). *La Pensée Sauvage*. Paris: Plon.

Levi-Strauss, C. (1996). *The Savage Mind*. Oxford: Oxford University Press.

Levon, E. (2016). Qualitative analysis of stance. In P. Baker and J. Egbert, eds., *Triangulating Methodological Approaches in Corpus-Linguistic Research*. London: Routledge, pp. 152–66.

Lischinsky, A. (2011). In times of crisis: A corpus approach to the construction of the global financial crisis in annual reports. *Critical Discourse Studies*, 8 (3): 153–68.

Liu, M. (2022). Stancetaking in Hong Kong political discourse. *Chinese Language and Discourse*, 13(1): 79–98.

Louw, B. (1993). Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In M. Baker, G. Francis and E. Tognini-Bonelli, eds., *Text and Technology: In Honour of John Sinclair*. Amsterdam: John Benjamins, pp. 157–76.

Love, R. (2020). *Overcoming Challenges in Corpus Construction: The Spoken British National Corpus 2014*. London: Routledge.

Love, R., Dembry, C., Hardie, A., Brezina, V. and McEnery, T. (2017). The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics*, 22(3): 319–44.

https://doi.org/10.1017/9781009168144 Published online by Cambridge University Press

Lutzky, U. (2021a). *The Discourse of Customer Service Tweets*. London: Bloomsbury.

Lutzky, U. (2021b). ‘You keep saying you are sorry’: Exploring the use of *sorry* in customer communication on Twitter. *Discourse, Context & Media*, 39. www.sciencedirect.com/science/article/pii/S2211695820300969#b0150.

Marchi, A. and Taylor, C. (2009). If on a winter’s night two researchers . . . a challenge to assumptions of soundness of interpretation. *Critical Approaches to Discourse Analysis Across Disciplines*, 3(1): 1–20.

Marchi, A. and Taylor, C. (2018). Introduction. In C. Taylor and A. Marchi, eds., *Corpus Approaches to Discourse: A Critical Review*. London: Routledge, pp. 1–15.

Markussen, T. (2005). Practising performativity. *European Journal of Women’s Studies*, 12(3): 329–44.

Mautner, G. (2005). The entrepreneurial university: A discursive profile of a higher education buzzword. *Critical Discourse Studies*, 2(2): 95–120.

Mautner, G. (2007). Mining large corpora for social information: The case of *elderly. Language in Society*, 36(1): 51–72.

Mautner, G. (2015). Checks and balances: How corpus linguistics can contribute to CDA. In R. Wodak and M. Meyer, eds., *Critical Discourse Studies*, 3rd ed. London: Sage, pp. 154–79.

Mautner, G. (2016). *Discourse and Management*. London: Palgrave.

Mautner, G. (2019). A research note on corpora and discourse: Points to ponder in research design. *Journal of Corpora and Discourse Studies*, 2: 2–13.

Mautner, G. (2022). What can a corpus tell us about discourse? In A. O’Keeffe and M. McCarthy, eds., *The Routledge Handbook of Corpus Linguistics*, 2nd ed. London: Routledge, pp. 250–62.

Mautner, G. and 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): 273–93.

Maxwell, J. A. (2013). *Qualitative Research Design: An Interactive Approach*. Thousand Oaks: Sage.

McEnery, T. and Baker, H. (2022). ‘A geography of names’: A genre analysis of nationality-driven names for venereal disease in seventeenth-century England. In T. Hiltunen and I. Taavitsainen, eds., *Corpus Pragmatic Studies on the History of Medical Discourse*. Amsterdam: John Benjamins, pp. 23–48.

McEnery, T., Baker, H. and Dayrell, C. (2019). Working at the interface of hydrology and corpus linguistics: Using corpora to identify unrecorded droughts in nineteenth-century Britain. In J. Egbert and P. Baker, eds., *Using Corpus Methods to Triangulate Linguistic Analysis*. London: Routledge, pp. 52–84.

McEnery, T. and Brookes, G. (2022). Building a written corpus: What are the basics? In A. O'Keeffe and M. McCarthy, eds., *The Routledge Handbook of Corpus Linguistics*, 2nd ed. London: Routledge, pp. 35–47.

McEnery, T. and Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

McEnery, T. and Xiao, R. (2006). *Corpus-Based Language Studies: An Advanced Resource Book*. London: Routledge.

Nartey, M. and Mwinlaaru, I. N. (2019). Towards a decade of synergising corpus linguistics and critical discourse analysis: A meta-analysis. *Corpora*, 14(2): 203–35.

Partington, A. (2004). Utterly content in each other's company: Semantic prosody and semantic preference. *International Journal of Corpus Linguistics*, 9 (1): 131–56.

Partington, A., Duguid, A. and Taylor, C. (2013). *Patterns and Meanings in Discourse: Theory and Practice in Corpus-Assisted Discourse Studies (CADS)*. Amsterdam: John Benjamins.

Partington, A. and Marchi, A. (2015). Using corpora in discourse analysis. In D. Biber and R. Reppen, eds., *The Cambridge Handbook of English Corpus Linguistics*. Cambridge: Cambridge University Press, pp. 216–34.

Pei, J., Li, D. and Cheng, L. (2022). Media portrayal of hackers in *China Daily* and *The New York Times*: A corpus-based critical discourse analysis. *Discourse and Communication*, 16(5): 598–618.

Pérez-Paredes, P. and Mark, G. (2021). *Beyond Concordance Lines: Corpora in Language Education*. Amsterdam: John Benjamins.

Phillips, J. C. and Egbert, J. (2017). Advancing law and corpus linguistics: Importing principles and practices from survey and content-analysis methodologies to improve corpus design and analysis. *Brigham Young University Law Review*, 6: 1589–620.

Pollach, I. (2012). Taming textual data: The contribution of corpus linguistics to computer-aided text analysis. *Organizational Research Methods*, 15(2): 263–87.

Pomerantz, A. (1984). Agreeing and disagreeing with assessments: Some features of preferred/dispreferred turn shapes. In J. M. Atkinson and J. Heritage, eds., *Structures of Social Action: Studies in Conversation Analysis*. Cambridge: Cambridge University Press, pp. 57–101.

Potts, A. (2015). Filtering the flood: Semantic tagging as a method of identifying salient discourse topics in a large corpus of Hurricane Katrina reportage. In P. Baker and T. McEnery, eds., *Corpora and*

https://doi.org/10.1017/9781009168144 Published online by Cambridge University Press

Discourse Studies: Integrating Discourse and Corpora. Basingstoke: Palgrave, pp. 285–304.

Potts, A. and Baker, P. (2012). Does semantic tagging identify cultural change in British and American English? *International Journal of Corpus Linguistics*, 17(3): 295–324.

Räikkönen, J. (2022). Are 'we' European? *We* and *us* in British EU-related newspaper articles in 1975–2015. *Journal of Corpora and Discourse Studies*, 5(1): 1–25.

Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics*, 13(4): 519–49.

Rebechi, R. R. (2019). God, nation and family in the impeachment votes of Brazil's former President Dilma Rousseff: A corpus-based approach to discourse. *Journal of Corpora and Discourse Studies*, 2: 144–74.

Reppen, R. (2022). Building a corpus: What are key considerations? In A. O'Keeffe and M. McCarthy, eds., *The Routledge Handbook of Corpus Linguistics*, 2nd ed. London: Routledge, pp. 13–20.

Sanscartier, M. D. (2020). The craft attitude: Navigating mess in mixed methods research. *Journal of Mixed Methods Research*, 14(1): 47–62.

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

Scott, M. (2020). *WordSmith Tools* (Version 8) [Computer Software]. Stroud: Lexical Analysis Software.

Semino, E., Demjen, Z., Hardie, A., Payne, S. A. and Rayson, P. E. (2018). *Metaphor, Cancer and the End of Life: A Corpus-Based Study*. London: Routledge.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, J. (1999). A way with common words. In H. Hasselgård and S. Oksefjell, eds., *Out of Corpora: Studies in Honour of Stig Johansson*. Amsterdam: Rodopi, pp. 157–79.

Sinclair, J. (2003). *Reading Concordances: An Introduction*. London: Pearson Longman.

Solan, L. M. and Gales, T. (2017). Corpus linguistics as a tool in legal interpretation. *Brigham Young University Law Review*, 6: 1311–58.

Stubbs, M. (1983). *Discourse Analysis: The Sociolinguistic Analysis of Natural Language*. Chicago: University of Chicago Press.

Stubbs, M. (1997). Whorf's children: Critical comments on critical discourse analysis (CDA). In A. Ryan and A. Wray, eds., *Evolving Models of Language*. Clevedon: Multilingual Matters, pp. 100–16.

https://doi.org/10.1017/9781009168144 Published online by Cambridge University Press

Stubbs, M. (2001). *Words and Phrases: Corpus Studies of Lexical Semantics*. London: Blackwell.

Stubbs, M. and Gerbig, A. (1993). Human and inhuman geography: On the computer-assisted analysis of long texts. In M. Hoey, ed., *Data, Description, Discourse: Papers on the English Language in Honour of John McH Sinclair on His Sixtieth Birthday*. London: Harper Collins, pp. 64–85.

Subtirelu, N. C. and Baker, P. (2018). Corpus-based approaches. In J. Flowerdew and J. E. Richardson, eds., *The Routledge Handbook of Critical Discourse Studies*. London: Routledge, pp. 106–19.

Taylor, C. (2013). Searching for similarity using corpus-assisted discourse studies. *Corpora*, 8(1): 81–113.

Taylor, C. (2022). The affordances of metaphor for diachronic corpora & discourse analysis. *International Journal of Corpus Linguistics*, 27(4): 451–79.

Taylor, C. and Marchi, A. (eds.). (2018). *Corpus Approaches to Discourse: A Critical Review*. London: Routledge.

Thornborrow, J., Ekstrom, M. and Patrona, M. (2021). Discursive constructions of populism in opinion-based journalism: A comparative European study. *Discourse, Context & Media*, 44. https://doi.org/10.1016/j.dcm.2021.100542.

Tkacukova, T. (2015). A corpus-assisted study of the discourse marker 'well' as an indicator of institutional roles: Professional and lay use in court cases with litigants in person. *Corpora*, 10(2): 145–70.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

Tribble, C. (2010). What are concordances and how are they used? In A. O'Keeffe and M. McCarthy, eds., *The Routledge Handbook of Corpus Linguistics*. London: Routledge, pp. 167–83.

van Dijk, T. A. (1993). Principles of critical discourse analysis. *Discourse & Society*, 4(2): 249–83.

Winter, B. (2019). *Statistics for Linguists: An Introduction Using R*. London: Routledge.

Wodak, R., Kwon, W. and Clarke, I. (2011). 'Getting people on board': Discursive leadership for consensus building in team meetings. *Discourse & Society*, 22(5): 592–644.

Wodak, R. and Meyer, M. (eds.). (2015). *Methods of Critical Discourse Studies*, 3rd ed. London: Sage.

Wright, D. (2017). Using word n-grams to identify authors and idiolects: A corpus approach to a forensic linguistic problem. *International Journal of Corpus Linguistics*, 22(2): 212–41.

Wright, D., Robson, J., Murray-Edwards, H. and Braber, N. (2022). The pragmatic functions of ‘respect’ in lawyers’ courtroom discourse: A case study of Brexit hearings. *Journal of Pragmatics*, 187: 1–12.

Zottola, A. (2021). *Transgender Identities in the Press: A Corpus-Based Discourse Analysis*. London: Bloomsbury.

Zottola, A., Jones, L., Pilnick, A. Mullany, L., Boumon, W., Arcelus, J. (2021). Identifying coping strategies used by patients at a transgender health clinic through analysis of free-text autobiographical narratives. *Health Expectations*, 24(2): 719–27.

https://doi.org/10.1017/9781009168144 Published online by Cambridge University Press

Cambridge Elements

Corpus Linguistics

About the Series

Corpus Linguistics has grown to become part of the mainstream of Linguistics and Applied Linguistics, as well as being used as an adjunct to other forms of discourse analysis in a variety of fields. It continues to become increasingly complex, both in terms of the methods it uses and in relation to the theoretical concepts it engages with. The Cambridge Elements in Corpus Linguistics series has been designed to meet the needs of both students and researchers who need to keep up with this changing field. The series includes introductions to the main topic areas by experts in the field as well as accounts of the latest ideas and developments by leading researchers.

https://doi.org/10.1017/9781009168144 Published online by Cambridge University Press

Cambridge Elements

Corpus Linguistics

Elements in the Series

Multimodal News Analysis across Cultures
Helen Caple, Changpeng Huan and Monika Bednarek

Doing Linguistics with a Corpus: Methodological Considerations for the Everyday User
Jesse Egbert, Tove Larsson and Douglas Biber

Citations in Interdisciplinary Research Articles
Natalia Muguiro

Conducting Sentiment Analysis
Lei Lei and Dilin Liu

Natural Language Processing for Corpus Linguistics
Jonathan Dunn

The Impact of Everyday Language Change on the Practices of Visual Artists
Darryl Hocking

Analysing Language, Sex and Age in a Corpus of Patient Feedback: A Comparison of Approaches
Paul Baker and Gavin Brookes

Shaping Writing Grades: Collocation and Writing Context Effects
Lee McCallum and Philip Durrant

Corpus-Assisted Discourse Studies
Mathew Gillings, Gerlinde Mautner and Paul Baker

A full series listing is available at: www.cambridge.org/corpuslinguistics