BJM BRITISH JOURNAL of MANAGEMENT BRITISH ACADEMY OF MANAGEMENT

British Journal of Management, Vol. 35, 1667–1679 (2024)
DOI: 10.1111/1467-8551.12816

Taking the Road Less Travelled: How Corpus-Assisted Discourse Studies Can Enrich Qualitative Explorations of Large Textual Datasets

Mathew Gillings,[1] Mark Learmonth[2] and Gerlinde Mautner[1]

[1]Institute for English Business Communication, Vienna University of Economics and Business, Welthandelsplatz 1, Vienna, 1020, Austria [2]Department of Human Resource Management, Nottingham Business School, Nottingham Trent University, 50 Shakespeare Street, Nottingham, NG1 4FQ, UK
Corresponding author email: mark.learmonth@ntu.ac.uk

How might interpretivist qualitative researchers tackle large data sets consisting of millions or even billions of words? Corpus-assisted discourse studies (CADS) is the approach we explore here. Specifically designed for the analysis of voluminous textual data, it offers a recognized empirical approach for making sense of such data. But it does so within an epistemology that understands language to be central in shaping our understanding of the world around us, so that CADS can assist researchers in revealing the social dynamics of the text – including the ideology and power that is latent in many such corpora. Bringing together the training of applied linguists and a management scholar, we discuss the background to CADS and its differences from text-mining approaches such as topic modelling, which have been more widely used in management studies to date. Focusing on the needs of people who are new to the approach, we then offer a worked example to show CADS' potential in exploring a management-related corpus. Our paper concludes with a discussion of the strengths and weaknesses of the approach and its potential for future discursively orientated management research – especially in the context of the rise of 'big data'.

Introduction

This paper aims to highlight the potential of an approach called *corpus linguistics* – a long-established computer-assisted method that facilitates the identification of linguistic patterns in large-scale textual data (Egbert, Larsson and Biber, 2020; Pollach, 2012). When linked to discourse analysis, it allows researchers to study large corpora from a qualitative, discursive perspective. This combination has come to be known as *corpus-assisted discourse studies*, or CADS[1] (Baker, 2006; Gillings, Mautner and Baker, 2023; Hardt-Mautner, 1995; Partington, Duguid and Taylor, 2013; Stubbs, 1996; Stubbs and Gerbig, 1993). Examples of the sort of corpora that could be explored using this method include all the debates in the UK parliament over a 40-year period (Perren and Dannreuther, 2012); extensive collections of company annual reports (Rutherford, 2005); British government policy documents over a 20-year period (O'Reilly and Reed, 2010); or the British National Corpus, a publicly available corpus of millions of naturally occurring texts such as newspaper articles and transcripts of spoken interactions (Cornelissen, 2008). While now fully established and widely recognized within applied linguistics, CADS has, to date, enjoyed a less prominent profile in management studies. There are of course notable exceptions, discussed later, but it is clear that CADS is not yet the 'go-to' method for dealing with large corpora when investigating management issues. It is still, to paraphrase Robert Frost's famous poem, 'the road less travelled'.

[1] CADS is the acronym for *corpus-assisted discourse studies*, the name of the field that uses corpora and computers to study discourse. Increasingly, the acronym is also being used to refer to the method (e.g. 'a CADS approach…'). What is more, as will be evident from our own paper, the acronym 'CADS' has become so independent that writers also use it to stand in for 'corpusassisted discourse analysis'.

Part of the reason for this lack of recognition is the dominance of what are often called text-mining approaches, including, in particular, topic modelling (see e.g. Andreou, Harris and Philip, 2020; Guerreiro, Rita and Trigueiros, 2016; Hajli *et al.*, 2022; Illia, Sonpar and Bauer, 2014; Tonidandel *et al.*, 2022). Over the last decade, these approaches seem to have established themselves as a popular option for management scholars, or even as the obvious choice when analysing large corpora (see Ghobadian *et al.*, 2022 for a recent example of topic modelling in this journal). Exciting work has been produced within this area, but text mining is both theoretically and operationally very different from CADS. One key difference is that most text-mining approaches share an underlying quantitative orientation. For instance, Hickman *et al.* (2022, p. 114), in their recent paper on text mining in organizational research, are typical in emphasizing 'the statistical power of … analyses, and the validity of insights derived from text mining'. Such quantitative perspectives tend to downplay questions of meaning, power and how language shapes our understanding of the world around us. Another major difference is that CADS and text-mining approaches have contrasting ideas about how language and meaning works. Text mining typically assumes that words operate in a way that is 'independent of syntax, narrative, or location within the documents' (Mohr *et al.*, 2013; in Hannigan *et al.*, 2019, p. 589), whereas CADS retains the syntax, narrative, location, etc. – factors understood as central to an understanding of how meaning is constructed in texts. We believe that these two differences are key reasons why text mining is likely to be relatively unattractive for researchers who adopt a constructivist, discursive view of the nature of language and its relationships with social life.

Of course, quantitative approaches are appropriate for many forms of research. However, the dominance of text-mining methods might be taken to imply that they are the *only* approach to the analysis of large datasets. This dominance, along with the often quantitative assumptions and rhetoric that can surround the use of text-mining, might suggest to many qualitative researchers that it is inappropriate, perhaps even impossible, for them to engage systematically with large-scale unstructured textual datasets as *qualitative* researchers. After all, traditional forms of qualitative work typically rely on researchers coding the data, but coding is impractical for textual datasets that run into millions or billions of words. Hence, we discuss CADS, an approach that makes the interpretative-hermeneutic process more manageable. We believe that it represents a way forward for researchers who wish to analyse large corpora and still maintain the view that language not only reflects but also constructs social reality (Fairclough, 1992; Mautner, 2016; Poorhosseinzadeh and Strachan, 2021).

The ontological and epistemological foundations of CADS are constructivist and interpretivist - an understanding that has been built into its design from the start. As such, they are radically different from those of text mining, which are typically orientated towards the conventional assumptions of quantitative researchers. Thus, we make no attempt at direct comparisons between CADS and text mining. Even though some have argued that text mining can be useful in qualitative work (e.g. Jacobs and Tschötschel, 2019), CADS and text mining do not simply do things differently; they do different things.

The primary audience for our paper, then, is scholars from management studies who (i) need to deal with large corpora, (ii) wish to do so through discursively orientated textual analysis on a recognized empirical footing, and (iii) eschew the implicit ontology and epistemology typical of text-mining approaches. Over 10 years on from Pollach's (2012) ground-breaking work on corpus linguistics for organizational scholarship, our distinctive contribution is to show the benefit of using CADS for management scholars in today's 'big data' environment, while emphasizing its value in unveiling the discursively mediated ideology and power latent in much textual data (Baker *et al.*, 2008).

In view of the rise and further mushrooming of big data and artificial intelligence over the last 10 years, the need for more discursively sensitive and critical methods for analysing large corpora of voluminous data is becoming ever more salient. In a nutshell, the biggest asset of CADS is that it allows the analyst to switch back and forth between close-ups and wide-angle views of the data – to focus on search words and their immediate surroundings, while also having access to the full texts. Among other things, and of particular interest to us, this combined approach allows – and indeed encourages – more critical perspectives to emerge. Thus, by bringing insights from a branch of linguistics to the table, we are able to tap analytical resources not as readily available to other disciplines. For those who would like more detailed practical advice, see Gillings, Mautner and Baker (2023) and Baker (2023). Our paper is not a methodological primer guide, but an encouragement to engage with a new method that you might not have previously had on your radar at all.

We proceed as follows. The paper begins by offering a general introduction to the currently dominant approach to the analyses of large corpora, namely text mining (an umbrella term referring to several different approaches to the quantitative study of big data), and specifically we look at the type of text mining referred to as topic modelling. While the latter has many strengths, we argue that it has been designed primarily for quantitative research, and so has significant limitations for discursively orientated researchers. We then move on to CADS, with an introduction to the methodological toolkit. We illustrate what CADS can do by providing an extended example drawn from our own work – a diachronic analysis of a leading management journal. We conclude with a look at possible futures for the analysis of large corpora within management studies, particularly from a discourse-sensitive perspective.

Approaches to large datasets in management studies

Before the digital era took hold, social scientists working with large volumes of unstructured data faced a stark choice: either content themselves with datasets small enough to permit coding and analysis, or settle instead for basic computational methods and abandon the idea of qualitative inquiry altogether. While the former option typically raised questions about cherry-picking because of the small samples involved (Baker and Levon, 2015; Mautner, 2015; Plakoyiannaki and Budhwar, 2021), the latter prevented the analyst from engaging with the data in a thoughtful and context-sensitive manner. The former sacrificed breadth, and the latter, depth.

With digital technology now in full swing, today's researchers have a broader range of choices. These choices are more important than they used to be because researchers are increasingly confronted with a veritable 'deluge of user-generated content (…) on the Internet' (Schmiedel, Müller and vom Brocke, 2019, p. 943). One possibility for dealing with voluminous collections of naturally occurring texts is subjecting them to what is often referred to by the umbrella term *text mining* (Frawley, Piatetsky-Shapiro and Matheus, 1992; Fan *et al.*, 2006; Guerreiro, Rita and Trigueiros, 2016; Kobayashi *et al.*, 2018).

Within management research, the text-mining approach that now dominates the field is known as *topic modelling* (Blei, 2012; Brett, 2012; Graham, Weingart and Milligan, 2012; Hannigan *et al.*, 2019; Tonidandel *et al.*, 2022). It can be described as 'a computer-aided technique [that] utilizes algorithms to generate a list of latent topics from a "corpus" (i.e. a given set of textual documents)' (Ghobadian *et al.*, 2022, p. 400). Topic modelling clearly holds great promise as 'a statistical tool for discovering abstract topics in a collection of texts' (Ristilä and Elo, 2023, p. 1). Without the need for resource-intensive coding, it allows the researcher 'to derive recurring themes from text corpora' (Schmiedel, Müller and vom Brocke, 2019, p. 943). Topic modelling is a machine-learning algorithm that takes a collection of texts as its input, transforms it into a so-called 'bag-of-words' (Brookes and McEnery, 2019; Shadrova, 2021; Gillings and Hardie, 2023), traces word co-occurrence at the level of the document and generates a number of unlabelled 'topics', defined as sets of statistically associated words. The output is twofold: a list of topics, each represented by a set of co-occurring words, along with a so-called composition document, which shows the quantitative distribution of topics across documents. It is then up to the researcher to decide what is an appropriate label for each list of statistically associated words.

However, the process of labelling an output tends to change how we perceive the output. We no longer see it as just a string of connected words, but instead as something much larger – a full topic that incorporates words with a range of meanings, connotations, ideological baggage and so on. In effect, a new label changes the topic's ontological status; put more simply, it changes its very nature. The label and thus the nature are up for revision, of course, but that does not change the fact that an act of reification has taken place. Furthermore, the topics have to be not only identified and labelled, but also interpreted in terms of their social meaning in a given context. And that is something that only the human researcher can do. Indeed, in spite of topic modelling's aura of objectivity, it cannot do without human-led intervention and interpretation (as most advocates of topic modelling would themselves readily acknowledge).

The key attraction of topic modelling nevertheless is its ambition to transform qualitative data into material that can be subjected to quantitative analysis. Therefore, in the text-mining literature, research problems are typically framed with the interests of quantitative researchers in mind. In other words, many accounts strongly emphasize the statistical, probabilistic bases of the approach and speak to concerns such as validity, reproducibility, and the predictive potential of the results.

The jury is still out on how powerful topic modelling is in doing justice to rich and multi-layered qualitative data, and how compatible it might be with the theoretical underpinnings of discourse analysis. A relatively small number of researchers are upbeat about the method's potential in this regard, particularly as an approach to triangulation or verification of qualitative discourse studies (Jacobs and Tschötschel, 2019, p. 478). When used on its own, however, several caveats are in order, as pointed out recently by several researchers. Two key steps involved in applying the technique are a particular cause for concern. The number of topics that the user asks the algorithm to return has been described as 'rather arbitrary', while the labelling of topics has been critiqued as 'an intuitive process' (Jaworska and Nanda, 2018, p. 383).

Unsurprisingly therefore, Brookes and McEnery (2019, p. 18) deplore 'the present lack of an adequate theoretical underpinning of what a topic actually is', which in their view 'has given rise to ill-defined (and likely inconsistent) procedures of topic discovery that lack linguistic sensitivity and are, it seems, liable to make the false assumption that unordered and de-contextualised words can be mapped neatly onto propositional topics'. As a result, 'analyses cannot be replicated or tested easily by others' (Gillings and Hardie, 2023, p. 5).

There are wider critiques from a specifically linguistic point of view; interested readers might wish to consult Brookes and McEnery (2019), Shadrova (2021) and Gillings and Hardie (2023). Among the criticisms, the following two appear to loom particularly large. The first calls into question how topic modelling deals with co-occurrence. Words are considered to be related if they appear *anywhere* in the same text, no matter how far apart. As Shadrova (2021, p. 13) argues, 'counting words as co-occurring when they are some 200 pages apart is a stretch of the concept'. The second concern relates to the difficulty of identifying topics on the basis of so-called content words alone (e.g. adjectives and nouns) and without taking function words such as prepositions into account. After all, to take a simple example, in terms of interpreting the meaning of a text, surely what matters is not only whether the word *capitalism* occurs in it at all, but also whether the word is preceded by the preposition *for* or *against*.

Corpus-assisted discourse analysis

The method of analysis that we favour is corpus-assisted discourse analysis. It is a triangulated method, combining corpus linguistics on the one hand with discourse analysis on the other, in order to look at how language operates in a wide variety of social domains. Using large amounts of textual data (often stretching to millions or billions of words), the researcher can look at micro-linguistic choices within a much wider discoursal and societal context. It begins with quantitative corpus-level word patterning, but then zooms in on points of interest for qualitative interpretation.

Significantly, at no point in the analysis does the researcher need to cut ties with the original texts. Analysts can go back and forth between viewing large-scale patterns across the corpus as a whole and then focus on specific linguistic items within their context – and indeed go back to the original documents. All of this takes place in the same corpus analysis program, rather than the researcher having to move between different files and tools. In other words, the radical decontextualization that text mining inherently involves does not arise in the first place. CADS thus links micro-linguistic choices with context, not only on a text level, but also on organizational, political, and societal levels. All these methodological advantages align with the standard concerns of good practice in traditional discourse analysis, regardless of whether or not computer assistance is used. For instance, CADS allows the analyst to be sensitive to factors such as intended meaning (Gries and Slocum, 2017), intended audience (Lutzky, 2023), genre conventions (Flowerdew, 2005) and other aspects of linguistic and discursive structures.

Furthermore, whilst it may be true that CADS is not widely used in management studies, there are exceptions. Authors in this field have used the method or similar approaches to link discursive changes to wider political and social themes. For example, Perren and Dannreuther (2012) traced 'the shifts in the discursive constructs of the entrepreneur that underlie political practice' (2012, p. 603), using a corpus made up of debates in the UK parliament over a 40-year period. They argue that 'diachronically tracing … constructed meanings of the entrepreneur provides insight into the institutional norms operating at particular times' (2012, p. 607). A related critical argument about the discourse of entrepreneurship in the context of CADS, albeit not diachronic, is made by Mautner (2005).

In a similar kind of way, O'Reilly and Reed (2010) examined a corpus of UK government texts from the 12-year period between 1997 and 2008. They found a major shift towards leadership within policy documents, evidenced by a dramatic rise in the frequency of the keyword *leadership*. This finding enabled them to undertake an extensive and in-depth discursive analysis of what they call 'leaderism'; that is, 'both a set of emergent discourses about leadership and as a set of framing metaphors encapsulating ideas of the process of "leading change" in the public services' (2010, p. 960).

Most recently, Stenka and Jaworska (2019) drew on Bourdieu's theorization of dominant discourse as a form of implicit power. Through a corpus-assisted study of a large number of comment letters submitted in response to proposals by an accountancy regulator, the authors demonstrate how:

> social agents argue for and discursively legitimize what they believe to be socially desirable outcomes in accounting regulatory debates. … [Their investigation] also opens new avenues for research in other domains of accounting where taken-for-granted discursive notions are being 'made-up' and used as justification for the actions of the key agents occupying … [powerful positions], very often without the active participation of those in whose interests such actions are (supposedly) taken (2019, p. 2).

Several dedicated corpus linguistics programs[2] have been developed to conduct these analyses, offering a way

[2] Popular corpus analysis programs include AntConc (Anthony, 2022), #LancsBox (Brezina, McEnery and Wattam, 2015), Sketch Engine (Kilgarrif *et al.*, 2014), CQPweb (Hardie, 2012), WordSmith Tools (Scott, 2020) and Wmatrix (Rayson, 2008). Programs differ according to whether they are free or subscription-based, whether they are available online or via

to present textual data that makes it easy to analyse the text discursively. These programs provide the following four traditional techniques: wordlists (together with frequencies), concordance analysis, collocation analysis and keyword analysis. Below, particularly with those new to the method in mind, we explore each of the four techniques and offer some sample analyses that might typically be conducted.

An illustration: Social actors in academic management writing

In what follows, we briefly summarize a study that two of us – one a linguist and one a management scholar – conducted recently (Mautner and Learmonth, 2020) with a view to illustrating the benefits, along with the challenges, of a CADS approach. To explore discursive change over time, with a focus on labels for social actors (e.g. *manager*, *worker*, *CEO*, etc.) we compiled and studied a corpus consisting of all the papers published in a major management journal, *Administrative Science Quarterly*, from the journal's creation to the end of 2018. The full corpus consisted of nearly 16 million words. We also created six subcorpora of 10 years each, and of roughly equal length, to facilitate longitudinal comparison. The software we used was Sketch Engine, a program widely employed by corpus linguists (Kilgarriff *et al.*, 2014), although most advanced corpus analysis programs offer very similar features.

As is commonly recommended in methodological guides (Gillings, Mautner and Baker, 2023), we used Sketch Engine's wordlist function as an entry point to the data. This is where words are ranked according to frequency of occurrence in the dataset. In this case, we searched for nouns only (which is possible in Sketch Engine because it automatically assigns words to grammatical categories). Given our interest in social actors, we picked out terms from the list that denoted this category, and given our interest in change over time, we compiled a wordlist for each of the six subcorpora.

A number of results were immediately apparent. For instance, even though the corpus was homogenous in the sense that it consisted only of management research, the wordlists showed a surprising lack of continuity. In fact, *manager, employee* and *member* were the only three social-actor labels that were consistently among the 100 most frequent nouns across all six subcorpora. While *manager* and *employee* might be expected to be common in management discourse, the reason why *member* was salient was not quite so obvious (and we will return to the term later). Another surprise was that *CEO* did not appear in the corpus until the 1980s, but had become the most frequent social-actor term of all by the 2010s. *Administrator* and *worker*, on the other hand, both declined notably in frequency over the time span covered, and both disappeared entirely from the top 100 nouns.

In order to be able to compare frequencies across our subcorpora, we also used Sketch Engine's ability to generate normalized frequencies (i.e. per million words) for words of interest. For example, Figure 1 compares the change in frequency of *administrator* and *CEO*.

Of course, the quantitative evidence garnered from wordlists and frequency lists is provisional and not a conclusive 'result' in its own right. In terms of Figure 1, for example, we would not want to claim that there is a direct (let alone causal) relationship between the decline of *administrator* and the rise of *CEO*. But such findings are intriguing and provide a first, tentative map of where and how more detailed, discursively sensitive searches might proceed.

Another promising path into the data is the keyword feature. A word is identified as 'key' if it occurs more frequently in one corpus than in another. Thus, keyword analysis allows us to contrast two corpora in order to investigate what is salient in a dataset. Using this technique, we are able to find out which words are, in relative terms, 'over-' or 'under-used' in a particular corpus, which can give us clues as to distinctive perspectives and communicative strategies.

Again, taking an example from our corpus, we compared the two most distant subcorpora (1950s/60s and 2010s). It emerged that the social-actor terms distinctive of the 1950s/60s subcorpus were *administrator, foreman, bureaucrat, teacher* and *superintendent*, while in comparison the distinctive key social-actor terms in the 2010 subcorpus were *CEO, investor, returnee, acquirer* and *marketer*. One possible interpretation of this difference is that in the most recent papers, writers were apparently much more concerned with financial elites and the upper echelons of corporations than their predecessors were. In the 1950s and 60s, by contrast, scholars' focus seems to have been more on the shop floor and the public sector.

For the discourse analyst, these are compelling clues – clues worth following up elsewhere in the corpus, but also outside of it. After all, the significance of the *assisted* in *corpus-assisted* is not to be underestimated; the corpus is often the main focus, but to explore the reason behind particular findings we may need to seek relevant complementary data such as media reports, policy documents, surveys, and so on. In any event, it is likely to be important that these clues have been identified computationally rather than on the basis of the researcher's intuition alone, not least to avoid the accusation of 'cherry-picking' convenient data. And while intuition and creativity of course remain factors that are extremely valuable, an external corrective – such as the one provided by corpus linguistics – means that

download, whether they have pre-made corpora available within them, and whether the user can upload their own data.

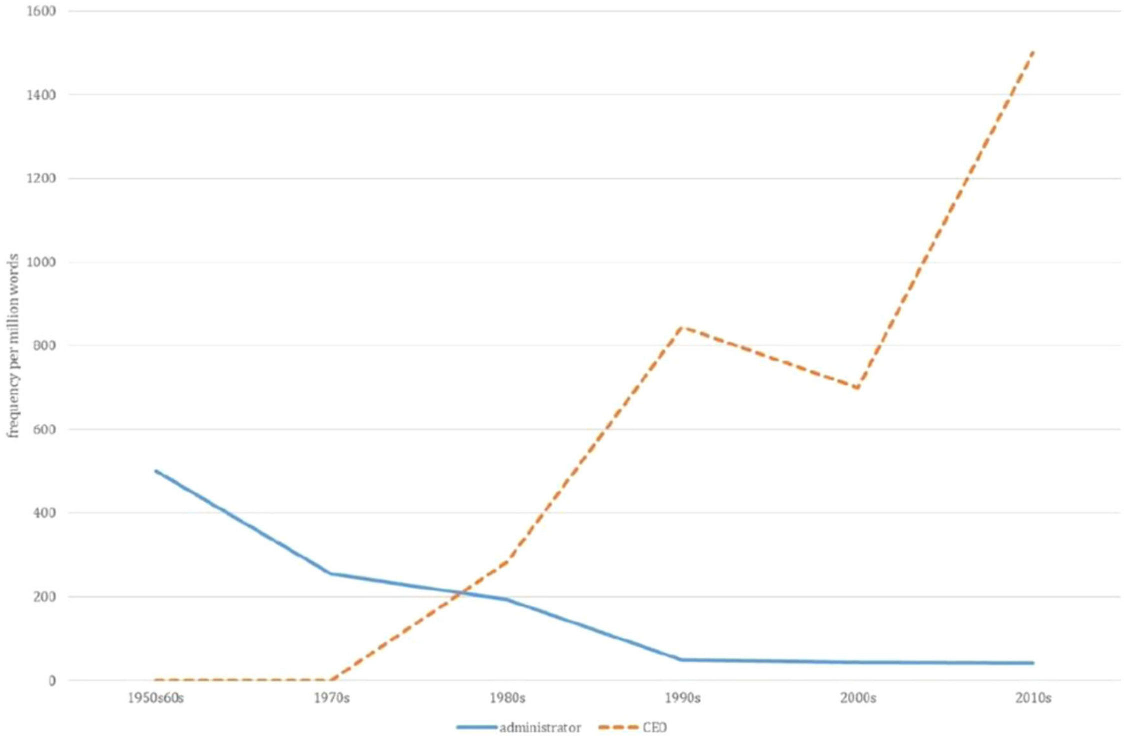
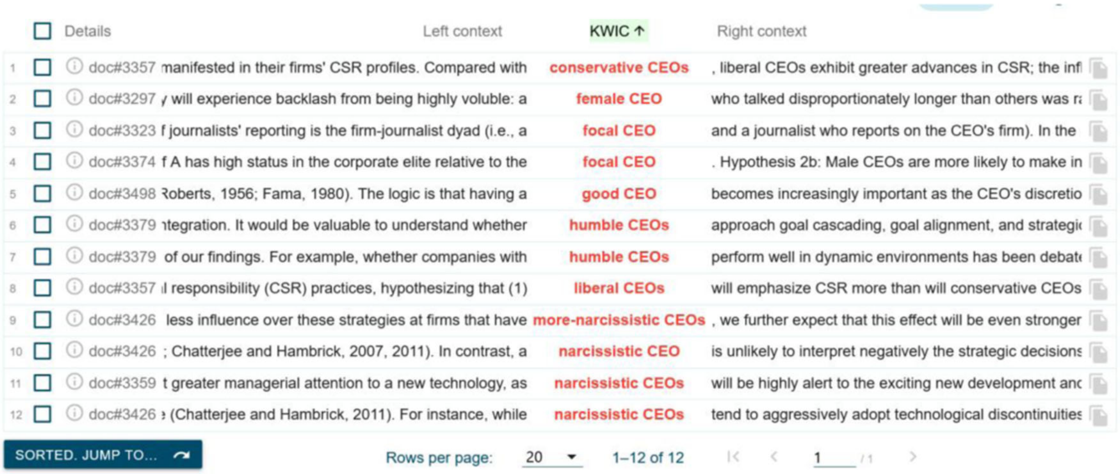


Figure 1. Frequency per million words of administrator *versus* CEO *across time (Mautner and Learmonth, 2020)*



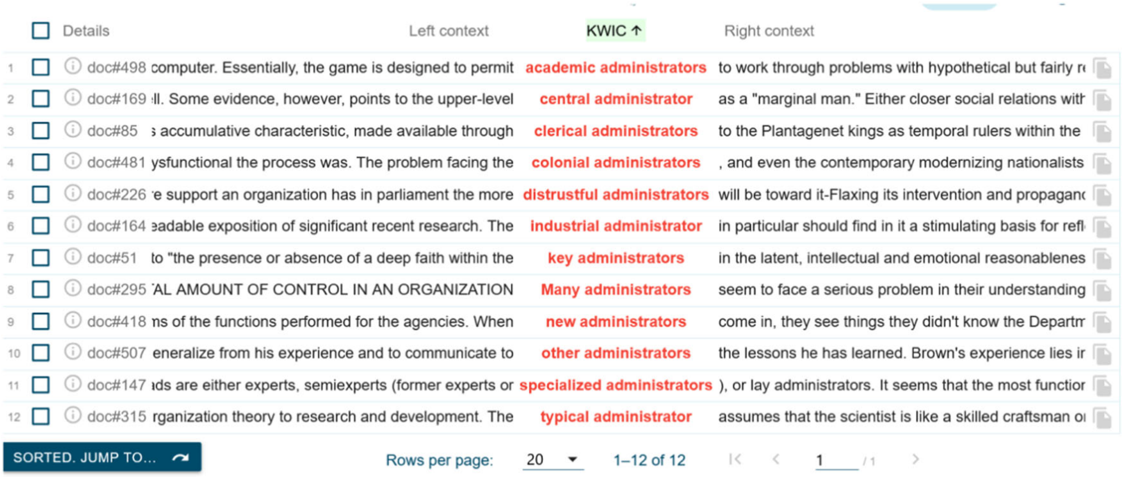
	□	Details	Left context	KWIC ↑	Right context
1	□	doc#3357	manifested in their firms' CSR profiles. Compared with	conservative CEOs	, liberal CEOs exhibit greater advances in CSR; the infl
2	□	doc#3297	y will experience backlash from being highly voluble: a	female CEO	who talked disproportionately longer than others was ra
3	□	doc#3323	f journalists' reporting is the firm-journalist dyad (i.e., a	focal CEO	and a journalist who reports on the CEO's firm). In the
4	□	doc#3374	f A has high status in the corporate elite relative to the	focal CEO	. Hypothesis 2b: Male CEOs are more likely to make in
5	□	doc#3498	Roberts, 1956; Fama, 1980). The logic is that having a	good CEO	becomes increasingly important as the CEO's discretio
6	□	doc#3379	ntegration. It would be valuable to understand whether	humble CEOs	approach goal cascading, goal alignment, and strategic
7	□	doc#3379	of our findings. For example, whether companies with	humble CEOs	perform well in dynamic environments has been debate
8	□	doc#3357	l responsibility (CSR) practices, hypothesizing that (1)	liberal CEOs	will emphasize CSR more than will conservative CEOs
9	□	doc#3426	less influence over these strategies at firms that have	more-narcissistic CEOs	, we further expect that this effect will be even stronger
10	□	doc#3426	; Chatterjee and Hambrick, 2007, 2011). In contrast, a	narcissistic CEO	is unlikely to interpret negatively the strategic decisions
11	□	doc#3359	t greater managerial attention to a new technology, as	narcissistic CEOs	will be highly alert to the exciting new development anc
12	□	doc#3426	e (Chatterjee and Hambrick, 2011). For instance, while	narcissistic CEOs	tend to aggressively adopt technological discontinuities

Figure 2. Screenshot from Sketch Engine, showing a random sample of 12 concordance lines for CEO(s) *preceded by an adjective (2010s subcorpus)*

researchers are less tempted to dig deeper only where they already know they are going to find something. Indeed, computationally identified themes may come as a complete surprise and thus open up interesting lines of inquiry that would not otherwise have occurred to the researcher. As a next step, to find out exactly how a particular keyword is used in context, we can look at so-called concordance lines, the feature to which we now turn.

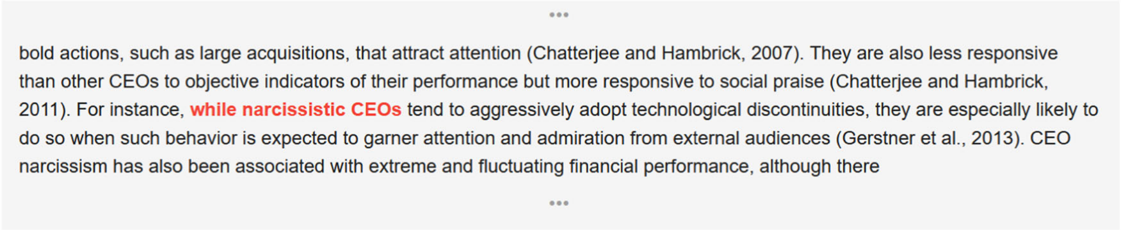
A concordance allows the researcher to view a search word (a ‘node’) alongside its linguistic co-text (i.e. a certain number of words stretching out to the left and right of the node). Searches can be made more focused by queries that specify a particular word class, such as noun or adjective. If you wanted to find out, for example, what qualities were typically ascribed to a particular type of social actor, you could look for occurrences of that actor, preceded by an adjective. Illustrations for *CEO* from the 2010s subcorpus and for *administrator* from the 1950s and 60s are given in Figures 2 and 3, respectively.

Reading such concordance lines takes some practice; as the lines are truncated, sometimes in the middle of a word, the investigator needs to get used to identifying



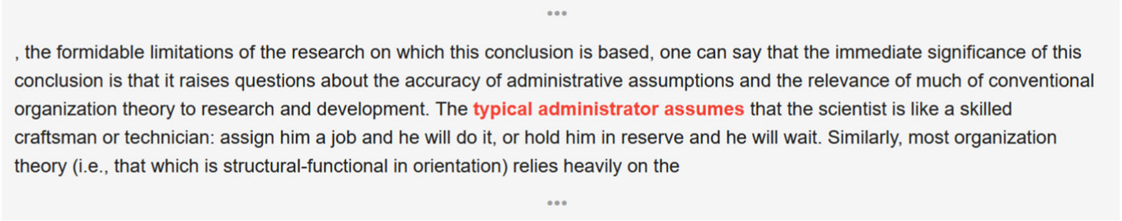
	□ Details	Left context	KWIC ↑	Right context
1	□ ⓘ doc#498	computer. Essentially, the game is designed to permit	**academic administrators**	to work through problems with hypothetical but fairly re
2	□ ⓘ doc#169	ıll. Some evidence, however, points to the upper-level	**central administrator**	as a "marginal man." Either closer social relations witł
3	□ ⓘ doc#85	; accumulative characteristic, made available through	**clerical administrators**	to the Plantagenet kings as temporal rulers within the
4	□ ⓘ doc#481	ysfunctional the process was. The problem facing the	**colonial administrators**	, and even the contemporary modernizing nationalists
5	□ ⓘ doc#226	'e support an organization has in parliament the more	**distrustful administrators**	will be toward it-Flaxing its intervention and propaganc
6	□ ⓘ doc#164	ɘadable exposition of significant recent research. The	**industrial administrator**	in particular should find in it a stimulating basis for refl
7	□ ⓘ doc#51	to "the presence or absence of a deep faith within the	**key administrators**	in the latent, intellectual and emotional reasonablenes
8	□ ⓘ doc#295	'AL AMOUNT OF CONTROL IN AN ORGANIZATION	**Many administrators**	seem to face a serious problem in their understanding
9	□ ⓘ doc#418	ns of the functions performed for the agencies. When	**new administrators**	come in, they see things they didn't know the Departm
10	□ ⓘ doc#507	eneralize from his experience and to communicate to	**other administrators**	the lessons he has learned. Brown's experience lies ir
11	□ ⓘ doc#147	ıds are either experts, semiexperts (former experts or	**specialized administrators**	), or lay administrators. It seems that the most functior
12	□ ⓘ doc#315	rganization theory to research and development. The	**typical administrator**	assumes that the scientist is like a skilled craftsman oı

Figure 3. Screenshot from Sketch Engine, showing a random sample of 12 concordance lines for administrator(s) *preceded by an adjective (1950s/60s subcorpus).*



bold actions, such as large acquisitions, that attract attention (Chatterjee and Hambrick, 2007). They are also less responsive than other CEOs to objective indicators of their performance but more responsive to social praise (Chatterjee and Hambrick, 2011). For instance, **while narcissistic CEOs** tend to aggressively adopt technological discontinuities, they are especially likely to do so when such behavior is expected to garner attention and admiration from external audiences (Gerstner et al., 2013). CEO narcissism has also been associated with extreme and fluctuating financial performance, although there

Figure 4. Concordance line 12 from Figure 2 expanded (output from Sketch Engine)



, the formidable limitations of the research on which this conclusion is based, one can say that the immediate significance of this conclusion is that it raises questions about the accuracy of administrative assumptions and the relevance of much of conventional organization theory to research and development. The **typical administrator assumes** that the scientist is like a skilled craftsman or technician: assign him a job and he will do it, or hold him in reserve and he will wait. Similarly, most organization theory (i.e., that which is structural-functional in orientation) relies heavily on the

Figure 5. Concordance line 12 from Figure 3 expanded (output from Sketch Engine)

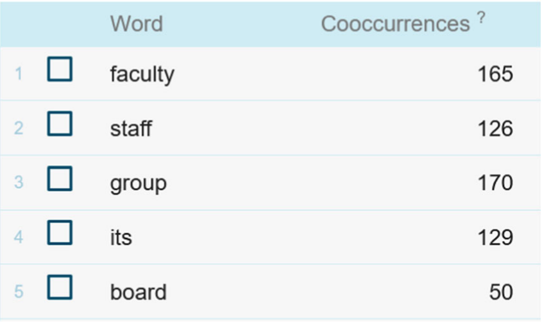
a suitable point at which to start reading in an analytical way. Once accustomed to the process though, researchers can use it to examine a large number of occurrences very efficiently. Indeed, there is great heuristic potential in viewing words in their context in this manner because the user can quickly read through examples and spot patterns that they would otherwise be likely to miss (Gillings and Mautner, 2024a; 2024b). Depending on the specific analysis, the researcher may choose to categorize these examples and thus gain a quantitative result, or they may simply read through them and learn more about a particular word's behaviour, as preparation for an in-depth qualitative analysis. Furthermore, should the single line given be insufficient to make sense of an example, the co-text can be expanded (in Sketch Engine's case, to five lines) in a mouse click. Figures 4 and 5 show what such longer extracts look like.

If an investigator needs a wider context still, they can expand the lines even further; the full text of the complete paper can be consulted because every line comes with a source identifier. In other words, the researcher can easily examine examples in detail in the same way that they would in a more traditional form of discourse analysis.

A final option offered by corpus analysis programs is the study of collocation; that is, examining the words that are statistically more likely to appear before and/or after a particular search word (i.e. more likely to appear than they would by chance). The concept of collocation in corpus linguistics is defined much more narrowly than word co-occurrence in topic modelling. In corpus linguistics, words are said to co-occur only if they appear within a narrow 'span', typically between one and 10 words to the left and right of the node word. The choice of span depends on the analyst's interests. Limiting the search to words directly next to the node is likely to focus on adjective–noun combinations, for example, whereas when words are included within a span of 10 say, the relationship begins to get more tenuous because the collocational magnetism diminishes with distance. In topic modelling, by contrast, words are said to co-occur as long as they appear within the same document.

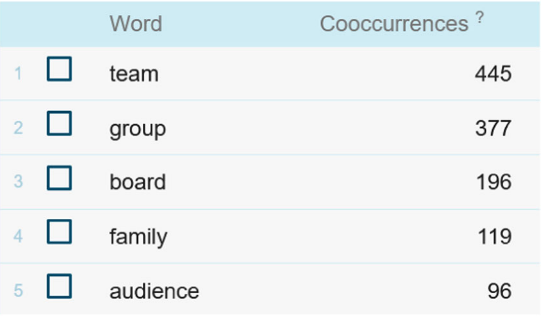
Especially those words that co-occur within a relatively narrow span can say a great deal about the word itself. Common word partnerships often have a story to tell about their wider social, political and historical context. Frequent collocates imbue a word with a particular 'semantic prosody'– that is, a positive or negative 'aura of meaning' (Louw, 1993, p. 157). For example, Learmonth and Morrell (2021) contrast the semantic aura of *leadership* and *management*. Arguing that the two terms are often used as apparent synonyms, they suggest that *leader* nevertheless 'strongly connotes images of greatness and goodness' (p. 2) (for instance, in its historic associations both with military heroism, and, at the same time, with figures like Martin Luther King Jnr or Mother Teresa, who are widely considered to have been 'good' (see p. 5). *Manager*, on the other hand, is 'a term which has considerably more negative and less prestigious associations' (p. 2) in that 'the connotations associated with "manager" now tend to conjure up a range of rather unheroic images linked to things like careerism, industrialism, authoritarianism, conflict, routine and bureaucracy' (p. 6). This sort of argument implies that we need to go beyond mere dictionary definitions when dealing with individual words, because the significance of the terms we use derives in large part 'from the social, cultural, and political environment in which they are embedded' (Mautner, 2005, p. 97). In this sense, meaning is not something that adheres to words independently. Rather, it drifts from one word to another, both affecting and being affected by neighbouring words through repeated usage in naturally occurring text and talk. Treating words in isolation is inadequate in that it misses these nuances.

Let us briefly return to our corpus and the puzzle over *member*, which we mentioned earlier. The frequency for *member* did not change over time, but we discovered that what *did* change was its collocates. In the 1950s/60s, the second strongest collocate (after *faculty*) was *staff* (as in the compound *staff member)*, whereas in subsequent corpora, the strongest collocate for *member* was *team* (as in *team member*; see Figures 6 and 7). Such findings open up interesting lines of more detailed inquiry into the discourses involved. In the case of *staff* versus *team*, it emerged that the two terms differed markedly in their semantic auras. Although they can be used to refer to the same group of people, the implications of using one or the other are very different. Arguably, *staff* emphasizes formal employment relations, whereas *team* appears to evoke harmony, interdependence and informality (Barker, 1999). The changing collocates of *member* might therefore suggest ideological change related to, for example, themes such as hierarchy and control, and the evolution of corporate rhetoric.



		Word	Cooccurrences ?
1	□	faculty	165
2	□	staff	126
3	□	group	170
4	□	its	129
5	□	board	50

Figure 6. The five most frequent collocates immediately before member(s) *in the 1950s/60s subcorpus (output from Sketch Engine)*



		Word	Cooccurrences ?
1	□	team	445
2	□	group	377
3	□	board	196
4	□	family	119
5	□	audience	96

Figure 7. The five most frequent collocates immediately before member(s) *in the 2000s subcorpus (output from Sketch Engine)*

On a practical note, when hovering over the question mark visible in the above two screenshots (located next to the word *co-occurrences*), Sketch Engine users are presented with an explanation of what is being shown in that particular column. Helpful tips like this, twinned with a responsive support team, make this corpus analysis tool intuitive and user-friendly. Whilst this is not to be taken for granted, the downside is that the service comes with a price tag, and so access to Sketch Engine and other commercial programs is restricted to those institutions and researchers who can afford it.

In concluding this section, it is important to emphasize that the four tools we have described – frequency, keyword analysis, collocations and concordances – can be used in any sequence and combined creatively. Frequency may be the starting point, but the researcher might want to go back to it at any later stage in the analysis – for example, when they have come across a particularly intriguing collocate. Similarly, Sketch Engine allows access to concordances straight from frequency lists, keywords and collocates, so that the analyst is less likely to lose sight of the subtleties of language use in the corpus being examined. CADS is an iterative

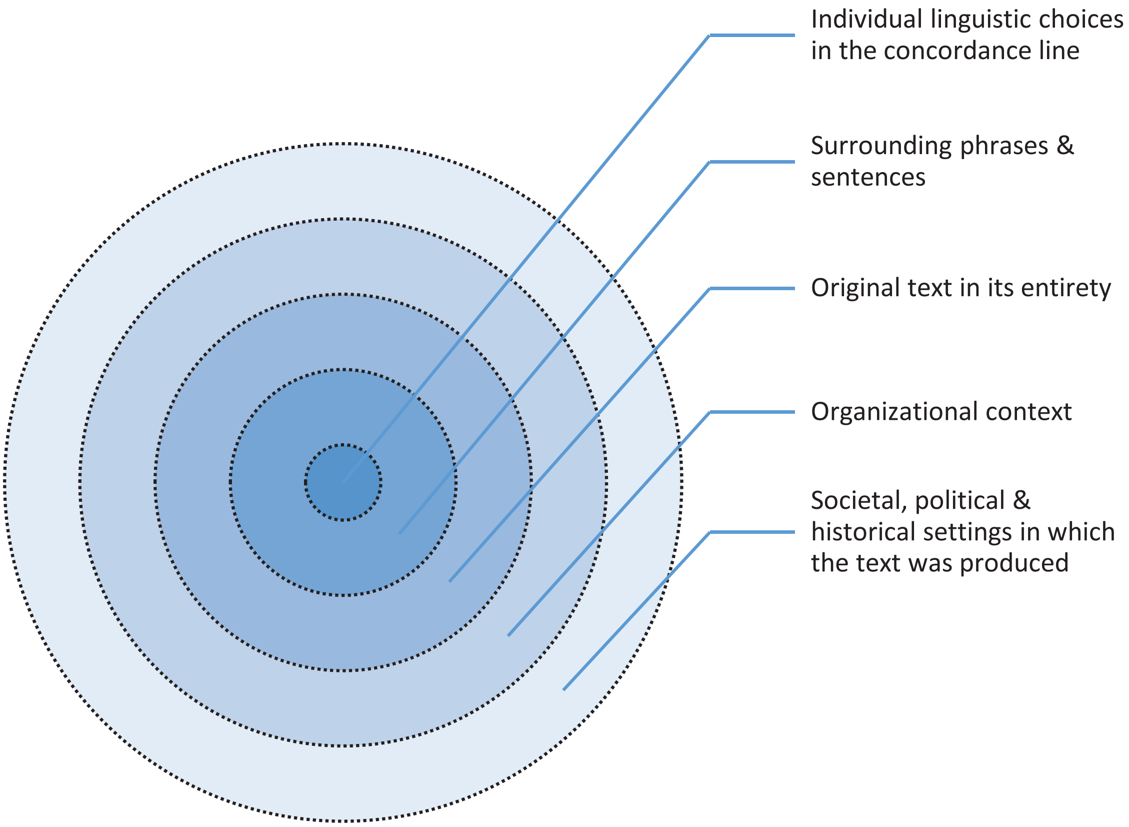

Individual linguistic choices in the concordance line
Surrounding phrases & sentences
Original text in its entirety
Organizational context
Societal, political & historical settings in which the text was produced

Figure 8. Embedded layers of analysis

Table 1. Key characteristics of CADS

CADS	
The main output of the analytical tool is …	a series of frequency lists, concordances, collocation tables, and lists of keywords. But not every project requires every technique.
Co-occurrence of words is defined as …	*collocation*. Words are said to collocate if they appear within a word span that is pre-defined by the researcher: typically, up to 10 words on either side of a search term.
Longer stretches of the original text are accessible …	directly from frequency lists, collocation tables and lists of keywords. Full texts are uploaded to the corpus analysis program, making individual concordance lines more easily interpretable.
The original structure of the texts …	is retained.
Coding …	is not a necessary prerequisite of the analysis, though researchers may choose to categorize findings after searches have been run.

process; it is not unsupervised statistical modelling, but a constant process where the researcher is responsible for utilizing each technique whenever necessary.

A certain degree of decontextualization is inevitable when looking at large datasets, it is true, but when using a corpus linguistics program, this effect is buffered by the permeability provided by the direct access between different levels of analysis. Figure 8 illustrates the interrelationships between these permeable layers. Individual linguistic choices, identified through concordance lines in the inner circle, can be linked to the surrounding phrases and sentences, the original text, the organizational context, and finally the political, societal and historical settings in which the text was produced. Table 1 summarizes the key characteristics of CADS.

Discussion and conclusion

In this paper, we hope to have established CADS as a valuable addition to the analytical repertoire of qualitative researchers who wish to work with large textual datasets in management and organizational studies. Major benefits include the following: (i) CADS allows both a macro perspective on the corpus as a whole and a micro view of nuanced linguistic detail; (ii) the researcher is able to zoom in and out of the data from within the same user interface; (iii) the researcher remains in close contact, throughout the investigation, with the full documents included in the corpus; (iv) CADS encourages sensitivity to the semantic aura of key terms without relying on manual coding; (v) it also enables researchers to avoid making *a priori* assumptions too early in the research process about what the data might be saying; (vi) CADS is relatively easy to use, in terms of the type of IT literacy required; and (vii) outputs are presented in a format congenial for qualitative analysis.

That said, why examine large corpora through a qualitative lens in the first place? The key point, over and above any purely academic interest, is this. There is no doubt that voluminous data, including so-called 'big data', is becoming hugely influential and that it is involved in every aspect of our public and private lives (Foster *et al.*, 2020). However, its more subtle yet pervasive influences are destined to remain under the radar if all we ever look at is content. It is therefore increasingly important for researchers to investigate not only *what* large corpora talk about but also *how* they do so, along with the resulting socio-political and ideological implications. CADS is well-placed to identify such nuances and assist the researcher in interpreting the linguistic and political choices shaping the discourses concerned. As a case in point, in our own illustrative example on management writing, we showed evidence of subtle shifts towards elite actors, corporate interests and a rhetoric apt to downplay organizational hierarchies and power differentials.

This sort of evidence has the potential to enrich the literature on – and add empirical detail to – a range of topics, including, for example, changes in corporate rhetoric and management fashions over time (Abrahamson, 1996; Cummings and Cummings, 2022; Oswick and Noon, 2014; Piazza and Abrahamson, 2020) or the interests being served when certain issues and debates gradually become hegemonic, while others are neglected (Alvesson and Blom, 2022; Learmonth and Morrell, 2019; Malhotra *et al.*, 2020; Le Pendeven, Bardon and Manigart, 2022). Related work has been done in applied linguistics – for example, in the context of the 2008 financial crisis (Lischinsky, 2011), climate change (Gillings and Dayrell, 2023) and corporate social responsibility (Fuoli, 2018). In these and other studies, CADS unveils linguistic patterning that might otherwise go unnoticed. It thus highlights the central role of discourse in constructing social structures, identities and processes.

However, in a paper such as this, which offers an alternative method in place of those that are generally used more widely, we would be remiss if we did not also engage in some critical reflection. While it is true that a corpus-assisted angle can offer unique insights that are likely to remain inaccessible through other methods, it is not the perfect fit for all investigations. And even when it is, certain caveats need to be kept in mind.

First, in spite of reminders that CADS is as much about 'discourse' as it is about 'corpus assistance' (Gillings, Mautner and Baker, 2023), our experience as reviewers suggests that it is not uncommon for researchers to privilege the latter over the former. Dazzled by the neatness of automated 'results', they can easily fall into precisely the trap that CADS was designed to avoid. Second, it is important to remember that the output of corpus-analysis programs is only ever a means to an end rather than an end in itself. From this perspective, the term *corpus-analysis programs* actually appears to be something of a misnomer. This is because the outputs we discussed (such as frequency lists, collocations and concordances) do not constitute an 'analysis' in and of themselves; they are only the basis on which a human researcher can then carry out the analysis. This is not a semantic quibble, but a substantive question of epistemology. Although there is always some fuzziness between how data is presented and how it speaks to the analyst, it is ultimately for the researcher to add to knowledge, not for the computer to do so. Third, queries using individual search words must not stop at that level, tempting though this may be. In fact, CADS does not fully come into its own unless the text surrounding the search term in question is examined qualitatively and with an eye to the discourse(s) manifested through collocational patterns. Fourth, standard corpus-analytical software is not (yet) set up to deal with visual material. Promising developments have been set in train – witness Bednarek and Caple's (2014) work on 'corpus-assisted *multimodal* discourse analysis', for example – but these are still to become widely and commercially available. For the time being, the 'CA' part of CADS is restricted to the textual level in the narrow sense, which means that exploring other semiotic modes is possible only as part of the 'DS' component of the undertaking.

Finally, in practical terms, studying corpora is best done with academic training in linguistics – about genre conventions, grammar and semantics, among other things. When a qualitative research design involves large volumes of textual data, it therefore makes sense for management scholars and linguists to work together, as has been the case for us in this paper. Yet the benefits of such interdisciplinary collaboration go far beyond the opportunities for tapping complementary expertise. Researchers socialized into different research routines can help each other out, certainly. But cooperating across disciplinary divides also has more substantive implications, for the development of theory as much as for the refining of methods. And perhaps most significantly, a perspective inspired by another discipline assists researchers not only in finding different answers, but in asking different questions, moving, for example, from an interest in recurring themes to an interest in recurring patterns of usage and nuances of meaning.

It is in that spirit that we encourage scholars from management and organization studies to use CADS in order to look at large-scale textual data through a discursive lens. They may find that 'the road less travelled' does indeed make 'all the difference' in terms of enriching management and organization studies, with both new theoretical and practical insights.

References

Abrahamson, E. (1996). 'Management fashion', *The Academy of Management Review*, **21**, pp. 254–285.

Alvesson, M. and M. Blom (2022). 'The hegemonic ambiguity of big concepts in organization studies', *Human Relations*, **75**, pp. 58–86.

Andreou, P. C., T. Harris and D. Philip (2020). 'Measuring firms' market orientation using textual analysis of 10-K filings', *British Management Journal*, **31**, pp. 872–895.

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

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

Baker, P. (2023). *Using Corpora in Discourse Analysis*. London: Continuum.

Baker, P., C. Gabrielatos, M. KhosraviNik, M. Krzyżanowski, T. McEnery and R. Wodak (2008). 'A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press', *Discourse and Society*, **19**, pp. 273–306.

Baker, P. and E. Levon (2015). 'Picking the right cherries? A comparison of corpus-based and qualitative analyses of news articles about masculinity', *Discourse and Communication*, **9**, pp. 221–236.

Barker, J. R. (1999). *The Discipline of Teamwork: Participation and Concertive Control*. Thousand Oaks, CA: SAGE.

Bednarek, M. and H. Caple (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**, pp. 135–158.

Blei, D. (2012). 'Probabilistic topic models: surveying a suite of algorithms that offer a solution to managing large document archives', *Communications of the ACM*, **55**, pp. 77–84.

Brett, M. (2012). 'Topic modeling: a basic introduction', *Journal of Digital Humanities*, **2**, pp. 12–16.

Brezina, V., T. McEnery and S. Wattam (2015). 'Collocations in context: a new perspective on collocation networks', *International Journal of Corpus Linguistics*, **20**, pp. 139–173.

Brookes, G. and T. McEnery (2019). 'The utility of topic modelling for discourse studies: a critical evaluation', *Discourse Studies*, **21**, pp. 3–21.

Cornelissen, J. P. (2008). 'Metonymy in language about organizations: a corpus-based study of company names', *Journal of Management Studies*, **45**, pp. 79–99.

Cummings, T. G. and C. Cummings (2022). 'Language and the evolution of academic fields: the case of organization studies', *Academy of Management Learning and Education*, **21**, pp. 598–623.

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

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

Fan, W., L. Wallace, S. Rich and Z. Zhang (2006). 'Tapping the power of text mining', *Communications of the ACM*, **49**, pp. 76–82.

Flowerdew, L. (2005). 'An integration of corpus-based and genre-based approaches to text analysis in EAP/ESP: countering criticisms against corpus-based methodologies', *English for Specific Purposes*, **24**, pp. 321–332.

Foster, I., R. Ghani, R. S. Jarmin, F. Kreuter and J. Lane (eds) (2020). *Big Data and Social Science: Data Science Methods and Tools for Research and Practice*, 2nd edn. Oxon: Routledge.

Frawley, W., G. Piatetsky-Shapiro and C. Matheus (1992). 'Knowledge discovery in databases: an overview', *Al Magazine*, **13**, pp. 57–70.

Fuoli, M. (2018). 'Building a trustworthy corporate identity: a corpus-based analysis of stance in annual and corporate social responsibility reports', *Applied Linguistics*, **39**, pp. 846–885.

Ghobadian, A., T. Han, X. Zhang, N. O'Regan, C. Troise, S. Bresciani and V. Narayanan (2022). 'COVID-19 pandemic: the interplay between firm disruption and managerial attention focus', *British Journal of Management*, **33**, pp. 390–409.

Gillings, M. and C. Dayrell (2023). 'Climate change in the UK press: examining discourse fluctuation over time', *Applied Linguistics*, **amad007**, pp. 1–24.

Gillings, M. and A. Hardie (2023). 'The interpretation of topic models for scholarly analysis: an evaluation and critique of current practice', *Digital Scholarship in the Humanities*, **38**, pp. 530–543.

Gillings, M., G. Mautner and P. Baker (2023). *Corpus-Assisted Discourse Studies*. Cambridge: Cambridge University Press.

Gillings, M. and G. Mautner. (2024a). 'Concordancing for CADS: Practical challenges and theoretical implications', *International Journal of Corpus Linguistics*, **29**, pp. 34–58.

Gillings, M. and G. Mautner. (2024b). '*Concordance lines: what are they and how can they be used to explore representation?*' In F. Heritage and C. Taylor (eds.), Analysing Representation: A Corpus and Discourse Textbook. Routledge.

Graham, S., S. Weingart and I. Milligan (2012). Getting Started with Topic Modeling and Mallet. Retrieved from: https://programminghistorian.org/en/lessons/topic-modeling-and-mallet (accessed 17 May 2023).

Gries, S. Th. and B. G. Slocum. (2017). 'Ordinary meaning and corpus linguistics', *BYU Law Review*, **6**, 1417.

Guerreiro, J., P. Rita and D. Trigueiros (2016). 'A text mining-based review of cause-related marketing literature', *Journal of Business Ethics*, **39**, pp. 111–128.

Hajli, N., U. Saeed, M. Tajvidi and F. Shirazi (2022). 'Social bots and the spread of disinformation in social media: the challenges of artificial intelligence', *British Journal of Management*, **33**, pp. 1238–1253.

Hannigan, T. R., R. Haans, K. Vakili and H. Tchalian (2019). 'Topic modeling in management research: rendering new theory from textual data', *Academy of Management Annals*, **13**, pp. 586–632.

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

Hardt-Mautner, G. (1995). '*Only Connect: Critical Discourse Analysis and Corpus Linguistics*', UCREL Technical Paper 6. Lancaster: University of Lancaster.

Hickman, L., S. Thapa, L. Tay, M. Cao and P. Srinivasan (2022). 'Text preprocessing for text mining in organizational research: review and recommendations', *Organizational Research Methods*, **25**, pp. 114–146.

Illia, L., K. Sonpar and M. W. Bauer (2014). 'Applying co-occurrence text analysis with ALCESTE to studies of impression management', *British Management Journal*, **25**, pp. 352–372.

Jacobs, T. and R. Tschötschel (2019). 'Topic models meet discourse analysis: a quantitative tool for a qualitative approach', *International Journal of Social Research Methodology*, **22**, pp. 469–485.

Jaworska, S. and A. Nanda (2018). 'Doing well by talking good: a topic modelling-assisted discourse study of corporate social responsibility', *Applied Linguistics*, **39** pp. 373–399.

Kilgarriff, A., V. Baisa, J. Bušta, V. Kovář, J. Michelfeit, P. Rychlý and S. Suchomel (2014). 'The Sketch Engine ten years on', *Lexicography*, **1**, pp. 7–36.

Kobayashi, V. B., S. T. Mol, H. A. Berkers, G. Kismihok and D. N. Den Hartog (2018). 'Text mining in organizational research', *Organizational Research Methods*, **21**, pp. 766–799.

Learmonth, M. and K. Morrell (2019). *Critical Perspectives on Leadership: The Language of Corporate Power*. New York, NY: Routledge.

Learmonth, M. and K. Morrell (2021). '"Leadership" as a project: neoliberalism and the proliferation of "leaders"', *Organization Theory*, **2**, pp. 1–19.

Le Pendeven, B., T. Bardon and S. Manigart (2022). 'Explaining academic interest in crowdfunding as a research topic', *British Journal of Management*, **33**, pp. 9–25.

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**, pp. 153–168.

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*, pp. 157–176. Amsterdam: John Benjamins.

Lutzky, U. (2023). '"Doesn't really answer my question…": exploring customer service interactions on Twitter', *International Journal of Business Communication*, **61**(1), pp. 92–114. https://doi.org/10.1177/23294884231200247

Malhotra, M., C. Zietsma, T. Morris and M. Smets (2020). 'Handling resistance to change when societal and workplace logics conflict', *Administrative Science Quarterly*, **66**, pp. 475–520.

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

Mautner, G. (2015). 'Checks and balances: how corpus linguistics can contribute to CDA'. In R. Wodak and M. Meyer (eds), *Methods of Critical Discourse Analysis*, pp. 154–179. London: SAGE.

Mautner, G. (2016). *Discourse and Management: Critical Perspectives through the Language Lens*. London: Palgrave Macmillan.

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

O'Reilly, D. and M. Reed (2010). 'Leaderism: an evolution in managerialism in UK public service reform', *Public Administration*, **88**, pp. 960–978.

Oswick, C. and M. Noon (2014). 'Discourses of diversity, equality and inclusion: trenchant formulations or transient fashions?', *British Journal of Management*, **25**, pp. 23–39.

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

Perren, L. and C. Dannreuther (2012). 'Political signification of the entrepreneur: temporal analysis of constructs, agency and reification', *International Small Business Journal*, **31**, pp. 603–628.

Piazza, A. and E. Abrahamson (2020). 'Fads and fashions in management practices: taking stock and looking forward', *International Journal of Management Reviews*, **22**, pp. 264–286.

Plakoyiannaki, E. and P. Budhwar (2021). 'From convention to alternatives: rethinking qualitative research in management scholarship', *British Journal of Management*, **32**, pp. 3–6.

Pollach, I. (2012). 'Taming textual data: the contribution of corpus linguistics to computer-aided text analysis', *Organizational Research Methods*, **15**, pp. 263–287.

Poorhosseinzadeh, M. and G. Strachan (2021). 'Straightjackets of male domination in senior positions: revisiting Acker's 'ideal worker' and the construction of the 'ideal executive"', *British Management Journal*, **32**, pp. 1421–1439.

Rayson, P. (2008). 'From key words to key semantic domains', *International Journal of Corpus Linguistics*, **13**, pp. 519–549.

Ristilä, A. and K. Elo (2023). 'Observing political and societal changes in Finnish parliamentary speech data, 1980–2010, with topic modelling', *Parliaments, Estates and Representation*, **43**, pp. 149–176.

Rutherford, B. A. (2005). 'Genre analysis of corporate annual report narratives: a corpus linguistics-based approach', *Journal of Business Communication*, **42**, pp. 349–378.

Schmiedel, T., O. Müller and J. vom Brocke (2019). 'Topic modelling as a strategy of inquiry in organizational research: a tutorial with an application example on organizational culture', *Organizational Research Methods*, **22**, pp. 941–968.

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

Shadrova, A. (2021). 'Topic models do not model topics: epistemological remarks and steps towards best practices', *Journal of Data Mining and Digital Humanities*, https://doi.org/10.46298/jdmdh.7595

Stenka, R. and S. Jaworska (2019). 'The use of made-up users', *Accounting, Organizations and Society*, **78**, pp. 1–17.

Stubbs, M. (1996). *Text and Corpus Analysis: Computer-Assisted Studies of Language and Culture*. Oxford: Blackwell.

Stubbs, M. and A. Gerbig (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*, pp. 64–85. London: Harper Collins.

Tonidandel, S., K. M. Summerville, W. A. Gentry and S. F. Young (2022). 'Using structural topic modeling to gain insight into challenges faced by leaders', *The Leadership Quarterly*, **33**, pp. 1–20.

Mathew Gillings is an Assistant Professor at the Vienna University of Economics and Business. Before moving to Vienna, he completed his PhD in Linguistics at Lancaster University. His research interests include the study of corporate wrongdoing, deception, and politeness in the workplace. He is a co-author of *Corpus-Assisted Discourse Studies* (CUP, 2023) and the author of *Corpus Approaches to Deception Detection* (Routledge, 2024).

Mark Learmonth is a Professor of Organization Studies at Nottingham Trent University. Prior to joining NTU, Mark worked at Durham University and the Universities of Nottingham and York. Before becoming an academic, he spent almost 17 years as an administrator in UK health care. His most recent book (with Kevin Morrell) is *Critical Perspectives on Leadership* (Routledge, 2019).

Gerlinde Mautner is a Professor of English Business Communication at the Vienna University of Economics and Business. Gerlinde was among the first to explore the opportunities and challenges involved in combining corpus linguistics and discourse analysis, and she continues to apply it to her own research. She is the author of *Language and the Market Society* (Routledge, 2010) and *Discourse and Management* (Palgrave, 2016).