Cardiff University Press
Gwasg Prifysgol Caerdydd

JCaDS
JOURNAL OF CORPORA AND DISCOURSE STUDIES 2023, 6(1):53–60
E-ISSN 2515-0251

BOOK REVIEW: GILLINGS, M., MAUTNER, G. & BAKER, P. (2023) *CORPUS-ASSISTED DISCOURSE STUDIES*. CAMBRIDGE ELEMENTS.

ALAN PARTINGTON
UNIVERSITY OF BOLOGNA / UNIVERSITÀ DI BOLOGNA

CITATION

Partington, A. (2023). Book Review: Gillings, M., Mautner, G. & Baker, P. (2023) *Corpus-Assisted Discourse Studies.* Cambridge Elements. *Journal of Corpora and Discourse Studies*, *6*(1):53–60

CONTACT

Alan Partington, Via S. Caterina da Siena, 5; 44121 Ferrara, Italy.
partington.alan@gmail.com

DOI

10.18573/jcads.109

ORCID

0000-0001-7881-80621

ISSUE DOI

10.18573/jcads.v6

Review accepted 2023-07-16

Book Review: Gillings, M., Mautner, G. & Baker, P. (2023) *Corpus-Assisted Discourse Studies.* Cambridge Elements.

Alan Partington
University of Bologna

Cambridge Elements, according to the publisher, is a series of quick, concise (approximately 100 pages) 'publishing solution for researchers and readers in the fields of academic publishing and scholarly communication [which] provide comprehensive coverage of the key topics in disciplines spanning the arts and sciences'.[1] One of the series of Cambridge Elements, to which the publication under review is a recent addition, is dedicated to Corpus Linguistics, with Susan Hunston as series editor. It is clear that brevity and comprehensiveness is a trade-off in all contexts. The Elements are somewhere between a handbook chapter on the field or topic in hand and an introductory volume, aimed at early-stage researchers.

The authors adopt a definition of 'discourse' as 'language as social practice' (p.1) which accords well with Halliday's (1985) widely-used description of discourse as 'language that is doing some job in some context' (p. 10). This is, however, followed by the startling assertion that '[w]hat all CADS [Corpus-Assisted Discourse Studies] projects have in common is that they have a social question [...] such as inequality, poverty, racism, or other social ills [...] at their centre rather than a purely linguistic one' (p. 1). This definition reflects the particular interest of the three authors, namely Critical Discourse Analysis (CDA), in which 'discourse' is a near synonym of social or political 'narrative', and which focuses largely on studying power differentials in society.

However, the term 'Corpus-Assisted Discourse Studies' was first coined at the initial Corpora and Discourse Conference at Camerino University, Italy, in 2002 (proceedings, Partington, Morley & Haarman [Eds.], 2004) during which corpus-assisted approaches to linguistics topics such as discourse organisation (Biber, Csomay, Jones & Keck), evaluation (Hunston), lexical priming and cohesion (Hoey, see Hoey 2005, 2017) and pragmatics in conversation analysis (Stenström) were discussed. *Ante litteram* CADS studies were conducted by Stubbs of, among other linguistics topics, modality (1996) and phraseology (2001), and see Biber, Conrad and Reppen (1998) on corpora and language use. There is now a growing body of corpus-assisted work on such linguistic and discourse fields as metaphor, literary stylistics, conversation analysis, (im)politeness, evaluation, semantic-discourse-evaluative prosody, evidentiality, recent language change, irony, humour and laughter-talk, translation and cross-language studies, and on register (academic and scientific writing, aviation, healthcare, legal discourse, even the language of movies). And,

[1] https://www.cambridge.org/core/publications/elements

in fact, the case study presented later in this Element (Section 5) is a linguistic investigation of a particular judicial register and into politeness.

In reality CADS is not confined by topic or the stance or school of the researcher/s, whether it be ontological and descriptivist, or normative and political. CADS is an *omnivore*: '[w]hat distinguishes CADS from traditional corpus research is the integration [when necessary] of additional information outside of the corpus during and after the linguistic analysis, namely through inductive, qualitative interpretation in order to uncover "non-obvious meaning" associated with a particular discourse type' (Skalicky, 2021, p. 591).

Much CADS work also concerns itself with language research methodology, including searching for absences, for similarities, using multiple data-sets, content analysis, the consequences of different ways of dividing data, multi-modal CADS, re-evaluating the qualitative-quantitative 'dichotomy', rather than having any particular social question as their core emphasis (Taylor & Marchi, 2018). CADS methodology is discussed in this Element in Section 2. The subset, then, of CADS which investigates social ills might best be defined as 'Corpus-Assisted Critical Discourse Analysis'.

With this caveat in mind, the Element remains a valuable guide to socially-focused CADS and many aspects can be generalized to all CADS.

Section 2 is in two parts. The first is a historical synopsis of the development of CADS, beginning with some of the influential early works, notably Hardt-Mautner (1995) and Stubbs (1997), moving on to a list of works in which a variety of social issues have been addressed, including those from fields such as business, law and healthcare. Sadly once again, little acknowledgement is made of the advances made in linguistic and discourse theory by CADS researchers.

The second part of the Section, entitled 'Flexible synergies' opens with a history of the pre-CL origins of CDA, a form of discourse studies which 'focuses on how discourse enacts ideology and power' (p.5). CDA had been heavily criticized for a lack of systematicity (Stubbs, 1997) and the potential danger of analysing only texts that 'attract our attention precisely because they seemed to provide the very evidence that the analysis was meant to uncover' (p. 6) and that, since 'the textual interpretations of critical linguists are politically rather than linguistically motivated [...] analysts find what they expect to find, whether absences or presences' (Stubbs, 1987, p. 111). Combining CDA analysis with CL methodologies seemed to offer a solution by enabling the systematic analysis of a larger number of texts. In particular, this combination can enable the researcher to employ mixed-method triangulations, that is, the use of more than one method to collect and analyse data, and to be able to shunt between statistical analysis and more traditional close-reading of segments of the corpus flagged as of interest by the quantitative overview (due perhaps to the particular frequency of an element, or infrequency or unexpectedness). It remains unclear however, as Stubbs (1997) underlines, the degree to which corpus-assistance alone reduces subjectivity or allows/forces researchers to escape the hermeneutic circle of confirmation bias/impulse. Corpora can be used simply to obtain more examples to support the starting hypothesis and give a veneer of greater objectivity, a danger mentioned in Section 6.

Section 3 contains an overview of the types of corpora which can be used to conduct CADS research. The important distinction is made between a corpus of language data and a database, which is grounded in the notion of representativeness, that is, corpora are compiled with the 'aim of representing a particular language variety' (p. 8). Although this definition is circular (in that the term being defined appears in the definition), the meaning is clear. A corpus is compiled in such a way as its contents are meant to be a balanced sample of texts from the universe of discourse under consideration though, as the authors stress, both the notion of what balance consists of and the decisions on which texts to include will involve a high degree of subjective choice. A randomly collected database of texts is more likely to be skewed by containing texts of a particular sort. An example might be as follows. If the researcher's objective was to compile a corpus of texts meant to be representative of corporate reports, the compilers might want to select a reasonably balanced number of reports from different types of companies (Alessi, 2016).

The variety of corpora discussed begins with two popular ready-made general or heterogeneric corpora (that is, corpora containing a variety of different discourse types and intended to be representative of the useful myth of 'general English') for British and American English, respectively, the British National Corpus 2014 (compiled at Lancaster University; see Love et al., 2017) and the Corpus of Contemporary American English (COCA) (Davies, 2008).

This section ends on the special issues posed by compiling a corpus of spoken English —not least time and funding—where decisions often need to be taken on what to include or exclude, for example, hesitations, pauses, overlaps, which for some research questions would be essential, for others a distraction. The discussion concludes by, tangentially, touching upon one of the original thorny issues of corpus linguistics, namely, how wise is it to familiarise oneself with the contents of a corpus? The authors note that: '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' (p. 12). One view is that the corpus should remain a black box so that researchers can analyse the data without preconceptions; another view is that this may be the case for lexico-grammatical analyses but, for discourse analysis, some prior knowledge of the discourse type can be useful in guiding the researchers to concentrate on interesting areas.

Section 4 takes the reader through some of the most important notions and methods of all corpus analysis, beginning with word and n-gram frequency, normalising frequency, especially when comparing different corpora, dispersion or clustering of items, tagging, concordancing and co-text, collocational analysis, including collocational networks and uncovering semantic prosodies, comparing corpora and keyword analysis.

Section 5 outlines a case study to help illustrate 'which tool is best suited at which point in the research process, and how the different tools are best orchestrated within a research design' (p. 38). The study is of a project investigating the particular linguistic register used for the speech act of expressing ‘respectful dissent’ by individual judges in judgements rendered by the UK Supreme Court. It introduces some of the tools available on SketchEngine,[2] including the ability to upload one’s own corpus and make use of the corpora that it provides for contrastive analysis, and the practice of ‘chain concordancing’, where an element noted in one concordance can lead the researcher to concordance anew that element, a recursive activity (the secret lying in knowing when to stop). The point of illustrating this case study is ‘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’ (p. 42), an excellent illustration of the omnivorous nature of CADS, and why the term ’corpus-*assisted*’ was deemed appropriate, as mentioned earlier.

Section 6 discusses some of the limitations and potential pitfalls of CADS research, beginning with the simple fact that it is not always the most useful or economical way of approaching a Research Question. The authors also caution that ‘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 & Mautner, 2023) and common collocates that simply do not seem to make sense. Such disappointing inconsistencies and contradictions cannot be argued out of existence’ (p. 48). Indeed, ‘inconvenient’ observations are often to be cherished as indications that one’s initial hypothesis needs to be refined. Partington, Duguid and Taylor (2013) argue for the ‘culture of the counterexample’ (331–338), that researchers should go out of their way to uncover inconvenient data (since, in the competitive spirit of the scientific method, if you don’t, someone else may well do).

The final Section, 7, is a collection of conclusions and broader considerations. The first is the gloomy acknowledgement that the take-up of CADS methodologies in other disciplines has been limited, despite the fact that *all* disciplines depend upon language for their research and dissemination. The authors suggest this may largely be due to the jealous natures of ‘disciplines’ ‘traditions’ and ‘schools’ (Mautner, 2016, 2019) and what I call the ‘Semmelweis’ effect,[3] the ‘not-discovered here’ attitude and ‘don’t tell us what to do’ syndrome. People working in different disciplines are used to employing their own ‘terminologies and rely on different ontological and epistemological assumptions: about something as basic as the nature and role of language, for instance’ (p. 50). However, part of the blame may also lie with the perception—openly expressed earlier in this Element —that CADS is only relevant to social issues and its occasional association with the so-

[2] https://www.sketchengine.eu/

[3] The term derives from the name of the Jewish Hungarian physician, Ignaz Semmelweis who, while working and researching in the Vienna General Hospital in 1847, discovered that childbed fever mortality rates fell ten-fold when doctors washed their hands between patients, especially after conducting an autopsy. Semmelweis's suggestion of simple hand-washing was frequently met with derision by fellow doctors who, after all, were the medical experts, Austrian gentlemen and Gentiles to boot. Semmelweis was finally vindicated by the gradual acceptance of the pathogen or germ theory of disease transmission in the second half of the 19thC. He is rightly remembered as the Saviour of Mothers.

called *critical* political stance, whereas, as stressed here, CADS methodologies are far more widely applicable.

The second reflection is the much happier one that, just like corpus linguistics, from a strictly Anglophone beginning, CADS is now being employed to study discourses in languages all around the world, including non-European languages, such as Arabic, Chinese and Japanese.

The final thought is that the sheer *mass* of data encountered in CADS can seem like a daunting *mess* of data. The authors recommend following what has been called the 'Mike Tyson principle', that is, have a plan, start out with as clear as possible a Research Question, whether this begins with a hypothesis or is inductive and data-driven, but be prepared for it to be mugged by the 'reality' of data observation, and be both flexible and honest enough to alter the plan. Be prepared too for underwhelming results, even after lengthy painstaking research; it would be a very strange methodology indeed that always returned novel, unexpected, exciting findings.

The upsides of the 'mess of mass' can be summarised as the benefits of *data overview*. Many discursive meanings are, as Baker (2006), puts it, incremental, in that they are built up and reinforced by being repeated and may therefore only become apparent to larger dataset analysis, especially when these are organized so as to capture repetition. The cumulative evidence provided by relatively large amounts of data can help expose the limits and liabilities of unassisted introspection. And, finally, the deliberate 'temporary alienation' in CADS of the analyst-observer-researcher from the object of research greatly enhances the prospects of the serendipitous discovery of non-obvious unforeseen information, so-called 'unknown unknowns', which can lead to entirely new avenues of research.

The Element ends with a handy list of some of the most frequently employed corpus analysis programs.

I would recommend this brief volume for use in an introductory course to CADS both for its practical tips and its reflections on the nature of research. It is complementary to and perhaps best read in combination with other recent overviews of Corpora and Discourse including Mautner (2019), Ancarno (2020) and Marchi and Taylor (2018), where several of the 'thorny issues' inherent in CADS work are discussed.

References

Alessi, G. M. (2016). Standardizing the language of corporate internal investigative reports: Linguistic perspectives on professional writing practices. In G. M. Alessi & G. Jacobs (Eds.), *The ins and outs of business and professional discourse research: Reflections on interacting with the workplace* (pp. 225–245). Basingstoke: Palgrave Macmillan. doi:10.1057/9781137507686_12

Ancarno, C. (2020). Corpus-assisted discourse studies. In A. De Fina & A. Georgakopoulou (Eds.), *The Cambridge handbook of discourse studies* (pp. 165–185). Cambridge: Cambridge University Press. doi:10.1017/9781108348195.009

Baker, P. (2006). *Using corpora in discourse analysis*. London: Continuum.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use.* Cambridge: Cambridge University Press.

Davies, Mark. (2008–). *The Corpus of Contemporary American English (COCA): One billion words, 1990-2019*. Retrieved from https://www.english-corpora.org/coca/

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

Halliday, M. A. K. (1985). *An introduction to functional grammar* (1st ed.). London: Edward Arnold.

Hardt-Mautner (1995). 'Only connect': Critical discourse studies and corpus linguistics. UCREL Technical Paper 6. Lancaster: Lancaster University.

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

Hoey, M. (2017). Cohesion and coherence in a content-specific corpus. In M. Pace-Sigge & K. J. Patterson (Eds.), *Lexical priming: Applications and advances* (pp. 3–40). Amsterdam: John Benjamins. doi:10.1075/scl.79.01hoe

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics, 22*(3), 319-344. doi:10.1075/ijcl.22.3.02lov

Marchi A., & Taylor, C. (2018). Introduction. In C. Taylor & A. Marchi (Eds.), *Corpus approaches to discourse: A critical review* (pp. 1–15). London: Routledge.

Mautner, G. (2016, July). Mad about method: Challenges and opportunities at the CDA/CL interface. Paper presented at the 3rd Corpora and Discourse Conference, University of Siena.

Mautner, G. (2019). A research note on corpora and discourse: Points to ponder in research design. *Journal of Corpora and Discourse Studies, 2*: 1–13. doi:10.18573/jcads.32

Partington, A., Morley, J., & Haarman, L. (Eds.). (2004). *Corpora and discourse*. Bern: Peter Lang.

Partington, A., Duguid, A., & Taylor, C. (2013). *Patterns and meanings in discourse: Theory and practice in corpus-assisted discourse studies (CADS)*. Amsterdam: John Benjamins. doi:10.1075/scl.55

Skalicky, S. (2021). Humorous and ironic discourse. In E. Friginal & J. A. Hardy (Eds.), *The Routledge handbook of corpus approaches to discourse analysis* (pp. 589–604). London: Routledge.

Stubbs, M. (1983). *Discourse analysis: The sociolinguistic analysis of natural language*. Oxford: Blackwell.

Stubbs, M. (1997). Whorf's children: Critical comments on critical discourse analysis (CDA). In A. Ryan & A. Wray (Eds.), *Evolving models of language* (pp. 100-116). Clevedon: Multilingual Matters.

Stubbs, M. (2001). *Words and phrases: Corpus studies of lexical semantics.* Oxford: Blackwell.