Cardiff University Press
Gwasg Prifysgol Caerdydd

JCaDS
JOURNAL OF CORPORA AND DISCOURSE STUDIES 2018, 1(1):1–7
E-ISSN 2515-0251

WELCOME TO THE FIRST ISSUE OF THE *JOURNAL OF CORPORA AND DISCOURSE STUDIES*

ALAN PARTINGTON

UNIVERSITY OF BOLOGNA

CITATION

Partington, A. (2018). Welcome to the first issue of the Journal of Corpora and Discourse Studies. *Journal of Corpora and Discourse Studies*, *1*(1):1–7

CONTACT

Alan Partington, Department of Interpreting and Translation, University of Bologna, Corso della Repubblica 136, Forlì, Italy. alanscott.partington@unibo.it

DOI

10.18573/jcads.19

ISSUE DOI

10.18573/jcads.v1i1

ORCID

—

Welcome to the first issue of the *Journal of Corpora and Discourse Studies*

Alan Partington
University of Bologna; Editor-in-Chief, JCaDS

The rationale behind the launch of the online open-access *Journal of Corpora and Discourse Studies* (*JCaDS*) is partly to meet the need for a discourse journal dedicated to research in which corpora play a significant role and partly to create a new corpus linguistics journal with a particular focus on discourse. Discourse is here defined as language in use as a vehicle of communication, as language doing things, as speakers and writers attempting to influence the beliefs and actions of their interlocutors using language.

But the rationale was also the realization that corpus-based, corpus-driven, corpus-assisted discourse analysis, corpus approaches to discourse, however we wish to name it, definable as 'that set of studies into the form and/or function of language as communicative discourse which incorporate the use of corpora' (Partington, Duguid and Taylor, 2013, p. 10), has for some considerable time matured into a field of study in its own right. It already has for instance, a biannual conference, and several volumes with 'corpora' and 'discourse' in their titles are on library bookshelves. As editors, then, we felt the time had come to provide a journal home, an on-line shop-window for the produce of this field of study.

We hesitate to use the term 'discipline', aware of Mautner's (2016) warning of how disciplines are prone to erect fences (perhaps necessarily so) by adopting specialized terminology and privileging certain methods and theories over others, and we wish corpora and discourse studies (henceforth CaDS) to remain as eclectic in language theory and welcoming of approaches and combinations of approaches as it has been until now, after all, '[a]n open mind is the best guide in linguistics, as in research in general and indeed in life itself' (Johansson, 1991, p. 6, quoted in Marchi, 2018). And it cannot be stressed early, often or strongly enough that approaches which incorporate the use of corpora and statistical techniques are not exclusive, in no way preclude or replace other approaches; they frequently marry well with, provide sustenance to, blend into and lead out of other types of approaches, and ways of collecting data (e.g. fieldwork, interviews, etc.; see Friginal and Hardy, 2014),[1] which is why CaDS is both particularly interdisciplinary and can be adopted in and adapted to so many other fields of study.

1 Interviews, questionnaires and such are often considered classic forms of qualitative research approaches. But when large enough quantities of such data are collected, they become available for statistical analysis, in other words they can become objects of so-called quantitative research.

Partington (2018) Welcome to the first issue of the Journal of Corpora and Discourse Studies. DOI 10.18573/jcads.19

In fact we expect and welcome contributions which apply discourse and corpora approaches to a wide variety of studies, including discourse organization and marking, cohesion, semantics, pragmatics, evaluation, importance-signalling, lexical grammar, conversation analysis, politics and institutional discourses, sociolinguistics, class and race issues, politeness, psycholinguistics, lexical priming, applied linguistics, stylistics, media, history, law, education, healthcare, economics, business and finance, gender studies, sexuality studies, cross-cultural studies, translation studies: indeed to any discourse-structure and topic area where more or less natural language is used as the vehicle of communication. However, having said all this, any field of study will display certain general characteristics, will offer particular opportunities and demonstrate limitations, and we will outline here some of those regarding CaDS.

Why then is it often productive to incorporate corpora techniques into discourse analysis? Many of the virtues can be summarized in the notion 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 be non-obvious in a small collection of texts but become apparent to larger dataset analysis, especially when these are organized so as to capture repetition.

Closely related and just as fundamental is that, in common with corpus linguistics research in general, the cumulative evidence provided by relatively large amounts of data can help expose the limits and liabilities of unassisted introspection; limits long-known and cautioned against, from Francis Bacon (1620/1848, p. 345) who argues that the intellect, left to itself, ought always to be suspected[2] to Richard Feynman (1974) who stresses that 'the easiest person to fool' is yourself, all of which is demonstrated to particular effect and on a regular basis in CL and CaDS; it is much harder to fool the machine. In addition to introspections, we need to make inferences and generalisations from the linguistic trace (the texts) left by speakers and writers, to build a model of their language behaviour. This tends to require a good number of texts. Other approaches to discourse analysis are certainly also data-driven, but large amounts of data is somehow *different* data or, as Sinclair puts it 'the language looks rather different when you look at a lot of it at once' (1991, p. 100). Large amounts of data, for example, make it much more feasible to look for counterexamples to a hypothesis we might wish to test, either one of our own or that of other researchers. It becomes difficult to ignore a large number of counterexamples which therefore force us to refine and improve or even reject the starting hypothesis, an essential part of the scientific process.

2 The sentiment in full, in English translation from the Latin, is even more relevant to CaDS: 'The unassisted hand, and the understanding left to itself, possess but little power. Effects are produced by the means of instruments and helps, which the understanding requires no less than the hand. And as instruments either promote or regulate the motion of the hand, so those that are applied to the mind prompt or protect the understanding.'

But corpus research is more than just exposure to large amounts of data. Corpus techniques open up a number of opportunities by virtue of allowing discourse researchers to *recontextualise* their data, often in several ways. Just as in other sciences (e.g., astronomy or chemistry), in CL and CaDS the discourse data is *re*-ordered, *re*-presented to view, even *re*-created, permitting the investigator to analyse it at different levels of abstraction. For example, concordancing allows us to view discourse 'vertically', which often reveals otherwise unsuspected patterns of regular usage. N-grams (also known as clusters or lexical bundles) and concgram techniques can also uncover typical ways of saying things — unconscious or deliberate lexical primings (Hoey, 2005) — across many tokens of a particular discourse type. Many other tools — from humble tables to histograms, to box plots, to heat-maps to word and semantic clouds, to dispersion plots, to scattergrams, to (interlocking) lexical network maps — provide visual representations of a series of phenomena, from, *inter alia*, raw or normalised frequency, to distribution and potential grouping, to the strength of collocational attraction among sets of lexical items (Anthony, 2018). All of which demonstrates how, as Stubbs (1996, p. 92) puts it 'you cannot understand the world just by looking at it' (…just one way). One of the early criticisms of corpus linguistics was that it only handles decontextualised language, but corpus linguistics and CaDS decontextualises in order to recontextualise and reconstruct the object of study, the discourse type under investigation.

And of course the abstractions — the frequency lists, semantic clouds, scattergrams, concordances, and so on — are performed by an entity, the machine, which is not the eventual interpreter and has no intuitive, primed expectations (Hoey, 2005) and no ideological vested interest. It is these processes of recontextualisation and the deliberate 'temporary alienation' of the analyst-observer-researcher from the object of research, their voluntary relinquishing of control over the research process, that act as a catalyst for the serendipitous discovery of non-obvious unforeseen information, the so-called 'unknown unknowns' which can lead to entirely new avenues of research, sometimes so many it becomes a (learned) intuitive skill in itself choosing which to most profitably follow up.

A further virtue of big-data overview in the analysis of discourse is the inescapable realisation that quantitative approaches not only complement qualitative ones, but that statistical information is often *in itself* functional information, that is, information on how linguistic items are used. Or, better, have been used, since all corpora are in effect archives of past language use. A couple of illustrations, the first strictly linguistic. A concordance of the item `fraught with` in UK newspapers, by presenting numerous examples of use in context shows quite plainly that the item has a negative prosody and three distinct semantic preferences, co-occurring with items from the sets of danger, problems/difficulties and negative emotions. This numerical data is clearly also functional information on how the item is used. The second illustration, the discovery

that the expression `Arab world` is found with greater frequency in newspapers published in that world than in the UK newspapers[3] and is therefore not an outsider term, and that the template `Egypt is [negative superlative] in the Arab world` is frequently found in an Egyptian-based newspaper (as of 2013), is equally important functional information; in this case it provides us with sociopolitical context on the possibility of media criticism in that country.

CaDS, as an intellectual activity, has not of course developed in a technological or financial vacuum.[4] While Hardt-Mautner (1995), Stubbs (1996) and Krishnamurthy (1996) were using corpora to study discourse in the 1990s, it was still possible in 2000 for McEnery and Wilson to note that 'discourse analysis is [an] area where the "standard" corpora have been relatively little used' (2000, p. 114). But the growing ease and cheapness of data collection has led to an explosion in the compilation of ad hoc 'bespoke' corpora, compiled to investigate a particular research question, often several corpora to study a single one. This has led to three more of the substantial virtues of using corpora in discourse analysis.

First, the ability to compare and contrast language phenomenon across different text-types, perhaps (im)politeness behaviours in different on-line fora, or politically divergent media stances on important sociopolitical issues, and so on. Second, we already noted above that many meanings are created incrementally, built up over repetition in many texts of the same type. We might add to this that they may also be created *transdiscursively*, that is, meanings can be reinforced by being passed among several different discourse types. They may, for instance be launched in political speeches, interviews or briefings, reappear in mainstream media comment, be picked up on social media, then find their way back into the official media via various dedicated 'social media watch' programmes and then onto the next day's press review and news programmes. Corpora can help us track these transdiscursive evolutions. Finally, it is also now possible to collect language data quickly and cheaply either periodically or continuously over time, which means, by comparing and contrasting different moments of such corpora, we can track both changes in language use and developments in social or political issues over recent periods of time (Davies, 2009), a sub-field of CaDS known as modern diachronic corpus-assisted discourse studies (Partington, 2010). And it should be stressed here that the particular capacity of CaDS for comparing and contrasting can reveal similarities as well as differences. Many corpus tools are designed to highlight the latter, a bias Taylor (2013) does well to warn us against.

The main advantages of on-line publishing are the speed with which works become available to the scientific community, but also the removal of frustrating financial

3 It is also used as a section heading on the Al-Jazeera Arabic website.

4 Few intellectual or scientific enterprises ever do. Without, for instance, the refining and evolving cheapness of glass, modern science would just not have happened.

impediments to researcher access. Studies are likely to get read more quickly, by a wider audience and thus feed into the body of knowledge more thoroughly.

The concluding opportunity provided by using corpora in discourse analysis is delivering transparency, one of the fundamental pillars of scientific research, the way in which, if we like, science is kept honest. Corpora are, if nothing else, physical archives. In ideal circumstances, if the composition and the architecture of the corpus are made clear and if the searches are documented and retrievable, each step of the analysis can be replicated by other researchers (and para-replicated on other, similar datasets to ascertain whether the same phenomena occur there). We say 'ideal' because these procedures have not always been apparent or possible due to various constraints, which include copyright and limited publication space, but also simple reluctance to share one's data. In order to maximise the accessibility of research across disciplinary boundaries and to foster open and critical analysis, *JCaDS* places emphasis on the explicit and comprehensive documentation of discovery procedures, and encourages authors to publicly deposit underlying data and analytic code whenever possible. The journal is therefore devised so that researchers can, if they wish, upload the data they used in their published research for the benefit of the wider community and we, as the editors, invite and encourage contributors to take advantage of this facility.

References

Anthony, L. (2018). Visualisation in corpus-based discourse studies. In C. Taylor & A. Marchi (Eds.), *Corpus Approaches to Discourse* (pp. 197–224). London: Routledge.

Bacon, F. (1848). Novum Organum (W. Wood, Trans.). In B. Montagu (Ed.), *The Works of Francis Bacon, Lord Chancellor of England* (Vol. 3). Philadelphia, PA: Carey & Hart. (Original work published 1620).

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

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990–present). *International Journal of Corpus Linguistics*, *14* (2), 159–190. doi:10.1075/ijcl.14.2.02dav

Feynman, R. (1974, June). *Cargo cult science: some Remarks on Science, Pseudoscience, and Learning How to not Fool Yourself*. Commencement address given at the California Institute of Technology. Retrieved from http://calteches.library.caltech.edu/51/2/CargoCult.htm

Friginal, E., & Hardy, J. A. (2014). *Corpus-based Sociolinguistics: A Guide for Students.* New York & London: Routledge

Hardt-Mautner, G. (1995). *Only Connect: Critical Discourse Analysis and Corpus Linguistics* (UCREL Technical Paper 6). Lancaster: Lancaster University. Retrieved from http://ucrel.lancs.ac.uk/papers/techpaper/vol6.pdf.

Hoey, M. (2005). *Lexical Priming. A New Theory of Words and Language*. London: Routledge.

Krishnamurthy, R. (1996). Ethnic, racial and tribal: the language of racism? In C. R. Caldas-Coulthard & M. Coulthard (Eds.), *Texts and Practices: Readings in Critical Discourse Analysis* (pp. 129–149). London: Routledge.

Marchi, A. (2018). *Self-Reflexive Journalism: A Corpus Study of Journalistic Culture and Community in The Guardian*. London & New York: Routledge.

Mautner, G. (2016, September). *Rough Crossings and Safe Havens: on the Challenges of Interdisciplinary Discourse Studies*. Paper presented at the CADAAD 2016 conference, Catania.

McEnery, A., & Wilson, A. (2000). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Partington, A. (Ed.). (2010). Modern-diachronic Corpus-assisted Discourse Studies. *Corpora*, *5*(2) (special issue). doi:10.3366/cor.2010.0101

Partington, A., Duguid, A., & Taylor, C. (2013). *Patterns and Meanings in Discourse: Theory and Practice in Corpus-Assisted Discourse Studies (CADS)*. Amsterdam & Philadelphia, PA: John Benjamins. doi:10.1075/scl.55

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

Stubbs, M. (1996). *Text and Corpus Analysis*. Oxford: Blackwell.

Taylor, C. (2013). Searching for similarity using corpus-assisted discourse studies. *Corpora*, *8*(1), 81–113. doi:10.3366/cor.2013.0035