Corpus-assisted Discourse Studies (CADS): Good Practices and Potential Pitfalls

Alan Partington

Bologna University

I want to start by outlining some of the relatively well-known methodological and epistemological achievements of Corpus Linguistics. I’d like then to show how these both feed into but differentiate from the requirements and practices of corpus-assisted discourse studies, defined as the employment of corpus techniques to shed light on aspects of language used for communicative purposes or, put another way, to analyse how speakers (attempt to) influence the beliefs and behaviour of other people (Partington, Duguid & Taylor 2013).

CADS does not refer to a particular school or approach, but is an umbrella term of convenience. Indeed, the types of research it refers to are eclectic and pragmatic in the techniques they adopt given that they are goal-driven, that is, the aims of the research dictate the methodology. However, although a broad church, it does possess its own characteristics, methods, resources, practices and is subject to its own particular temptations and pitfalls.

By means of various case studies, I want to illustrate the added values of CADS to discourse study. It can supply an overview of large numbers of texts, and by shunting between statistical analyses, close reading and analysis types half-way between the two, CADS is able to look at language at different levels of abstraction. After all, ‘you cannot understand the world just by looking at it’ (Stubbs 1996: 92), and abstract representations of it need to be built and then tested. Indeed, far from being unable to take context into account (the most common accusation levelled at Corpus Linguistics), CADS contextualises, decontextualises and recontextualises language performance in a variety of ways according to research aims. It also highlights how statistical information, sometimes dismissed as ‘merely’ quantitative, is actually inherently also qualitative in nature. Corpus techniques greatly facilitate comparison among datasets and therefore among discourse types. They can, moreover, ensure analytical transparency and replicability (and para-replicability). And because parts of the analysis are conducted by the machine, they enable the human analyst to step outside the hermeneutic circle, to place some distance between the interpreter and the interpretation. Finally, they enable the researcher to test the validity of their observations, for instance, by searching for counterexamples (‘positive cherry-picking’).

Having said all this, the discourse analytical process is always guided by the analyst, and there are many parts of the process which a machine simply cannot tackle. This is why we prefer the term ‘corpus-assisted’ to alternatives such as ‘corpus-driven’ or ‘corpus-based’.

The aim is to show how CADS sits within the wider framework of scientific research methodology, what we might mean by scientific objectivity in discourse analysis and what counts as good (in the senses of both ‘useful’ and ‘honest’) practices and what practices are best avoided.

**References**

Partington, A., A. Duguid and C. Taylor. 2013. *Patterns and Meanings in Discourse: Theory and Practice in Corpus-assisted Discourse Studies*. Amsterdam & Philadelphia: John Benjamins.

Stubbs, M. 1996. Text and Corpus Analysis: Computer-assisted Studies of Language and Culture. Oxford: Blackwell.