7
Bridging the gap between corpus tools and discourse analysis

Caroline Clark

Abstract

This paper examines how quantitative analyses typical of Corpus Linguistics can combine with more detailed qualitative approaches normally associated with Discourse Studies to provide valuable insights into discourse types. This "dual" approach, known as Corpus-Assisted Discourse Studies (CADS), allows the researcher – whether scholar or student – to manoeuvre between very large quantities of text while making systematic reference to textual details which may otherwise be "lost" in the huge quantity of data. An example is presented of class work using two large newspaper corpora, which were analysed using the *Wordsmith Tools* software (Scott 2011). This analysis was the basis of active student contribution to the research. The implications of the findings are discussed, as are some of the long-standing and unresolved issues regarding the use of corpus tools in discourse analysis.

7.1. Introduction

Investigating and analysing texts for research purposes, or as student course-work, generally termed Discourse or Text Analysis, have benefited from recent advances in technology. The use of computer tools has expanded the possibilities for analysis, although the question of the compatibility of Corpus Linguistics (hereafter CL) and Discourse Analysis (hereafter DA) remains unresolved. From both theoretical and methodological points of view, the issue is often avoided rather than confronted. An outline of the two approaches to text could contribute to illustrating the potential areas of incompatibility which may arise from the perception that CL and DA appear to be mutually exclusive.

Advances in computer technology have provided two crucial opportunities: the possibility to process data at great speed on the one hand, and the capacity to store massive amounts of readily (and quickly) accessible information on the other. This has opened a new horizon for the language researcher, but has also

o discussion of some methodological issues, such as what has been "lost"
what has been "gained" by computer processing of texts.

Corpus Linguistics approach

The CL approach is typically based on a large number (often millions) of
ds which are processed using specialised computer software for various
cepts based on wordlists, keywords and concordancing, in order to study
uencies of usage and word/phrase use in context. The methodology is clearly
eficial for very large studies based on numerous texts, and the analysis
ied out is inevitably quantitative, that is, numerical results will be produced
d on various statistical formulae. Research questions are therefore more
y to be oriented towards certain areas, such as lexicography.

This methodology remains purely descriptive unless it is comparative, either
hronically (comparing texts at a specific point in time) or diachronically
nparing the same types of text over time). Despite processing large quantities
xt, and applying sensitive statistical measures, the results for a single text or
onogeneric corpus of texts are descriptive in that they describe themselves
are hence rather meaningless. On the other hand, this approach provides
researcher with the opportunity to compare various large corpora and it
is possibility which lends itself to diachronic analyses, where language
ges can be tracked over time.

The data produced, and the form in which they are produced, lend themselves
ductive reasoning. That is, initial specific observations and measures of the
sen texts are made with the aim of detecting patterns and regularities in
r to formulate some tentative hypotheses that can be explored. CL studies
nguage are found in two main areas: language structure and language use,
is, revealing the ways that structures occur, how and when they occur,
ell as the context and functions. The Corpus Linguist may, therefore, be
rested in syntactic and morphological changes in language, lexicography, or
uage for special purposes.

One of the criticisms addressed to corpus analysis in quantitative terms is
"this sort of methodology can count words, but it cannot interpret them"
ig 1989: 206), that is, it can produce numerous quantitative results, but it does
offer the possibility of interpreting these results. There is also the concern
important features of context of production will be lost (Partington *et al.*
: 3), including context and cotext. Above all, it must be remembered that
rpus is a quantity of text which, alone, can give no information about the
uage (Hunston 2002); it is simply the raw materials.

7.3. Discourse Analytic approach

The DA approach is necessarily confined to single or limited series of texts with reduced numbers of tokens since the time investment is massive with large quantities of data. This type of analysis is basically "qualitative": it means that it has less to do with instances, and more with grades, where the topic of interest is often the underlying social structures which may be played out in the text, and the tools and strategies used in communication,

While this type of analysis can be extremely sensitive and comprehensive, and, above all thorough, it is time and energy consuming, especially if dealing with unwieldy quantities of data (Wilson 1993: 6). The DA approach tends not to be comparative, except in limited cases, and often pursues a "limited" question, rather than providing a general overview. The analysis is usually deductive, or top-down, where a specific hypothesis may be tested by seeking confirmation or otherwise in a corpus of texts. Hence DA research will usually be concerned with single texts and qualitative readings of contextualised data (Hardt-Mautner 1995: 24). The researcher may have to limit the quantity of data by choosing just one "example", in the form of a single work, edition, programme etc. given the quantity of data. However, limiting the size of the corpora may both weaken the possibility of convincing findings, as well as raising objections as to researcher interference in selection.

Further, it is necessary to evaluate whether a single text be considered as representative of the "whole", and the extent to which any conclusions drawn can be applied to the whole. Such use could also be considered as a variation of the "observer's paradox"[1] (Labov 1972), that is, if the researcher must select text, then the text is automatically charged with a possibly unmerited importance and symbolic status as representative of similar texts.

7.4. Corpus-Assisted Discourse Studies

The interesting question is whether some aspects of DA, in particular, the investigation of underlying discursive patterns and structural features, can be explored by a CL approach, and whether some CL procedures such as a greater number of texts and the application of specialised software which permits patterns and recurrences to emerge, may be appropriate to DA research.

What happens when we bring CL and DA together? The Corpus-Assisted Discourse Studies (henceforth CADS) approach merges the quantitative and qualitative methods of CL and DA which complement each other as part of a multi-strategy design which imposes an empirical dimension to introspection

[1] Labov's *observer's paradox* refers to the paradoxical situation of a researcher having to observe how people speak/write when *not being observed.*

Partington and Duguid 2010). The chosen research is reinforced by gulation (Webb *et al.* 2006), whereby multiple methodologies are used, r than relying exclusively on a single approach.

he CADS approach allows the researcher to manage and analyse an vhelming amount of material (visual and aural, as well as textual). titative results and data (typical of CL), and large numbers of tokens, can alysed without dismissing the contextual and cotextual features. This is combined with the qualitative approach (typical of DA) where other more ed, and less immediately obvious, informational levels and aspects of the urse can be retrieved. There is increased potential to refer systematically pects of the discourse which may otherwise be neglected, or obscured by rge quantities of data. It may also be argued that "attention to cotext is mised by having whole texts [...] accessible, available for different analyses" ngton 2004: 3).

he aim of CADS is to investigate and compare features of various discourse and in particular the more occult meanings which may not be immediately ble on surface reading, by integrating DA with the tools and techniques . As a form of corpus-driven approach, the corpus serves as an empirical from which to extract data and detect linguistic phenomena without assumptions and expectations (Tognini-Bonelli 2001). Any conclusions or s are made exclusively on the basis of corpus observations.

he corpora are analysed by applying specialised software such as *smith Tools* to produce data in the form of word and keyword lists, which hen be analysed further, using concordancing tools, for prosodic features reference to cotextual and contextual features remains available in the al form.

ADS at work

illustrate how this fusion may be put into practice, two large corpora tish newspapers dated thirteen years apart[2] are used as an example. This part of a much larger project is based on the apparent increased usage ormal, spoken language in 2005 newspapers, mirroring the perceived idization"[3] of the British quality papers.

iBol 93 Corpora comprise all the words (100 million) in the *Guardian*, *The Times* and the *Telegraph* in the year 1993 and SiBol 05 (145 million words) the same papers in 2005. The of years was dictated by availability of material. The study was part of a project carried out arch groups at the Universities of Bologna and Siena, Italy.

idization refers to the style of newspaper presentation which is concerned with enter- nt rather than information, more visuals and less text, gossip and scandal, shock headlines, nerally, it is associated with a lowering of journalistic standards.

As part of their English course based on the language of newspapers, B2 level students enrolled on degree courses in Political Science were invited to participate in the following activities as a classroom project. The aim was two-fold: to acquaint them with natural language in use, and to explore research techniques.

While criticism of tabloidization of the press has been widespread, it was not at all clear where (or whether) evidence of this would be found in the papers. In other words, the research was based on a "hunch" and evidence supporting this hunch was sought in the available data. From the phenomenon selected for examination, it was clear that the study must be comparative: an analysis of the 2005 corpora alone is meaningless. It was also immediately clear that the sheer size of the corpora does not permit a closer reading, or acquainting oneself with the texts.

Given the inevitable time restrictions, one of the main considerations was that students should not spend time compiling and accessing the corpus or having to learn to use the specialised software. They were presented with previously compiled corpora, as well as selected and prepared material, mainly in the form of lists, which they commented on and discussed.

7.6. Drawing up wordlists

As a first step into the corpora, two wordlists (one for each year) were extracted, and made available to students as *excel* spreadsheets. These files allowed students to reflect that little useful data was forthcoming, except to note an almost 50 percent increase in words in newspapers thirteen years later, and that this information is irrelevant to evaluating the existence (or not) of the phenomena of tabloidization.

The need to compare the two corpora was evident, and a keyword analysis was proposed. That is, each corpus was compared with the other as reference corpus to show those words with a far greater relative frequency in one or other corpus. The resulting two files of 5,000 words each (1993 keywords, and 2005 keywords) essentially show the difference between the papers at a distance of thirteen years and become the backbone of the following activities.

Students were provided with previously compiled keyword lists as spreadsheets to acquaint themselves with the data and to note any patterns that may be evident. It was clear from the keywords that, as expected, they were "overloaded" with situation and time specific words, mainly the names of people and places which were news in one period and not the other, as well as "products" (such as *Viagra*, *chocolate* and *alcohol*) and "e-words" (*www*, *click*, and *blog*) which are found among the 2005 keywords.

7.7. Lemmatisation

It was immediately evident that the lists should be reformulated in such a way that the useful information "floats to the top"; the original wordlists, for example, needed to be grouped or lemmatised to make differences more meaningful and to help words emerge. Automatic lemmatising is an extremely valuable tool, without which the true value of the corpus would be lost, and the researcher's work frustrating and dispersive. Lemmatising involves joining related words to a headword, such as the grouping of the singular and plural forms of nouns, or all forms of verbs.

After lemmatising the wordlists[4] and re-compiling the keyword lists, the first evidence of conversational language started to emerge, and certain trends became clear. Pronouns (*he, she, it*, etc, grouped together), and contracted forms (*I'm, can't*, etc), previously obscured, were now evident among the keywords, and honorifics (*Mr, Mrs*, etc) were among the negative keywords.[5] These findings led to the hypothesis that language has changed towards a tendency for the informal and colloquial, intended as the avoidance of erudite and bureaucratic words; reference to first names and nicknames which hint at face-to-face discourse, contracted and elided forms, reflecting speech and the use of deictic devices, in particular first and second person pronouns.

These findings were not unexpected as the quality British press has been criticised for dumbing-down, or tabloidization. Having identified these elements, it was hypothesised that an increased frequency of usage might be present for other examples, such as taboo words, given the increased familiarity of the language.

7.8. Example: taboo words

Without lemmatising, taboo words are effectively "lost" in the data as the relative frequency of each single taboo word is very low, and none of them appear as a keyword before lemmatisation. In order to explore any shifts in usage it was necessary to recover all these words. Given the nature of the words, they also appear in duplicate form, that is, with and without substitutive asterisks (e.g. *shit* or *s****).

The first problem encountered was what should be classified as a taboo word.[6] A further problem was merely technical: the use of replacement asterisks, which are a notation not recognised by the *Wordsmith* program. This meant that the instances of taboo words could be neither quantified nor concordanced. To resolve the problem, the texts had to be run through a text replacement programme (such as *SearchAndReplace*, Nodesoft v2) with the command that three or more consecutive asterisks (***) corresponding to letters be replaced by an alpha-numerical code (in this case *kkk*) so that they become retrievable. It was found that, surprisingly, the percentage of asterisked taboo words was almost the same in both corpora, despite the increase in taboo words overall. It was expected that the usage of asterisks would have diminished as newspapers moved towards a more "familiar" language – although there is a wide discrepancy between papers. For example, *The Times* asterisked 73 percent of taboo words in 2005 compared with only 19 percent thirteen years earlier which is a reversal of the overall trend. That is, the use of a timid *** to avoid possible offence did not diminish although the total of explicit non-asterisked taboo words did.

The next question was: what constitutes a taboo word? It was decided that a taboo word should be considered as any word which appears with asterisks, and its non-asterisked counterpart. A lemma list was thus formed and the wordlists re-calculated, which led to taboo words appearing as a 2005 keyword. The relative frequency of taboo words, appearing explicitly or censored by asterisks, across all papers increased by 150 percent, thus confirming our hypothesis of a shift towards a more spoken style of language.

Further investigation showed that the increased relative frequency of taboo words varied greatly across papers, with the greatest increase of all found in *The Times*, which had by far the lowest frequency (7 pmw) in 1993, which increased by 470 per cent to 40 pmw in 2005. However, this is only part of the picture.

7.9. Qualitative results: concordancing

Having evaluated the quantifiable data, and having the satisfaction of being able to draw some empirical conclusions from the findings, it was necessary at this point to embark on a quality based research by investigating the context of usage using the *concordancing* features (Baker 2006) of *Wordsmith*, again with data provided in the form of *excel* files to study the collocations. The aim was to reveal the linguistic patterns in which taboo words were found, who uttered them, the context of utterance as well as the discourse practice of taboo language.

The *Wordsmith* data used was again prepared and made available in the form of *excel* files. Since all taboo words appear in the word and keyword lists as *kkk*, this "word" was *concordanced*, that is, searched in all the text files and displayed with surrounding text. Thus, we were able to see the *collocates* (words in the neighbourhood) of all taboo words, and the *clusters* in which these words

[4] A lemma list compiled by Someya (1998) was used as a basis for further additions.

[5] Negative keywords are words which occur significantly infrequently, that is, words which occur less than would be expected, and are key in the other keyword list.

[6] Swear words, or words and phrases that may be considered by readers to be offensive, or inappropriate to a particular context.

appeared. We now had access to the context of taboo word usage, that is, we were able to describe how a word was in fact used, which may be quite different from dictionary entries.

Concordancing *kkk* made it possible to see that taboo words were found almost entirely within quotes (which also had to be marked up for the same reasons and using the same procedure as for taboo words) – a finding that was not unexpected. It was also apparent that taboo words can be found predominantly in the *sports* and *arts* pages of the quality papers. Using the *collocates* function (which shows the words most commonly found to the left and right of *kkk*), it can be seen that taboo words collocate highly with *you* and *I*, and the verbs *said*, *going* and *give*. It was interesting to observe that the first lexical words after the above verbs were *ball*, *money*, *people* and *black*. While the first suggests the sports pages, it is less likely that the second indicates the financial pages.

Concordancing can be refined further by looking at the *clusters*, that is, the patterns of repeating phraseology where *kkk* appears within a set number of words. It was also possible to see how taboo words were distributed across the various papers (using the *plot* function).

Returning to the *concordance* feature, chunks of text, from one line to the entire story, can be retrieved for further study. Thus, the data is reduced to a manageable size for qualitative study. The indications here do not outline exciting findings but rather they help the researcher identify where to look to study the phenomenon further.

7.10. CADS in the Classroom

The use of corpora for research purposes typically comprises three steps: *compiling* the corpus, *accessing* it using specialised software, and then *analysing* the results. Using large pre-compiled corpora with language students is not without problems, in particular regarding the sheer mechanics of dealing with it, the complexity of the software, and, unfortunately, generalised student diffidence towards it. The procedure entails time and specific knowledge, but it is not out of reach of students, especially as a class project for higher level students in smaller groups, using a pre-compiled corpus.

With higher-level students, the advantages of data-driven learning (Johns 1991) are two-fold. Johns argues that using corpora in the classroom gives students access to naturally occurring language in a form that text books are unable to imitate, and encourages them to test hypotheses about how words and phrases are used (see also Kilgariff 2009). Further, according to Gabrielatos (2005) concordances can also become a form of "condensed reading". Students are invited to observe the lists, draw conclusions and formulate hypotheses based on these observations. The intention is not to acquaint the student with

the language of newspapers, but the language of modern life and to raise awareness of the changing nature of language use in the social reality.

7.11. Conclusion

The aim of this short paper was two-fold; first, a brief outline of how two different approaches to large quantities of text, the quantitative and the qualitative need not be regarded as distinct and mutually exclusive. Instead, they can complement each other when analysing very large corpora, making it possible to access deeper strata of meaning. Within this context, was the aim of illustrating how this type of study can also be carried out successfully with students after having taken steps to make the resulting data more accessible. Having the data presented to students in previously prepared excel files, allowed them to focus on certain aspects without being overwhelmed by the sheer volume of material. In this way, they were able to invest more time (and less frustration) reflecting on research methodology and hypotheses within the context of naturally occurring language, which, in turn, had to be evaluated within a qualitative, or discourse analytical, frame.

7.12. References

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

Billig M. (1988), "Methodology and Scholarship in Understanding Ideological Explanation". In Antaki C. (ed.), *Analysing Everyday Explanation: A Casebook of Methods*, London, Sage, 206.

Gabrielatos C. (2005), "Corpora and language teaching: Just a fling, or wedding bells?", *TESL-EJ*, 8 (4), 1-37.

Hardt-Mautner G. (1995), "Only Connect", *Critical discourse analysis and corpus linguistics*, University of Lancaster. http://ucrel.lancs.ac.uk/papers/techpaper/vol6.pdf (22/01/2015).

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

Johns T. (1991), "From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning". In Johns T., King P. (eds.), *Classroom concordancing*, *ELR Journal* 4, University of Birmingham, 27-45.

Kilgariff A. (2009), "Corpora in the classroom without scaring the students", *Proceedings 18th International Symposium on English Teaching*, Taipei.

Labov W. (1972), *Sociolinguistic Patterns*, Oxford, Blackwell.

Partington A., Duguid A. (2010), *Modern Diachronic Corpus-Assisted Discourse Studies (MD-CADS)*, Pisa, Felici Editori.

Partington A., Morley J., Haarman L. (eds.) (2004), *Corpora and Discourse*, Bern, Peter Lang.

Scott M. (2011), *Wordsmith Tools,* (version 5.0), Oxford, Oxford University Press.

SearchAndReplace, Nodesoft V2.6.1.1 Online. http://www.nodesoft.com/.

Someya Y. (1998), E-*Lemma.* http://www.lexically.net/downloads/e_lemma.zip (06/2008).

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

Webb E., Campbell D., Schwartz R., Sechrest L. (2000), *Unobtrusive measures* (revised edition), London, Sage.

Wilson A., McEnery T. (1994), "Teaching and Language Corpora", Technical Report Department of Modern English Language and Linguistics, University of Lancaster.