# Corpus-Assisted Discourse Studies (CADS)

## Video 2 transcript

Full resource, see: <https://www.ncrm.ac.uk/resources/online/all/?id=20855>

In this video we look at some of the corpus linguistic techniques you might use as a way in to your CADS analysis. First, here’s a quick recap from video one on some of the entry points into analysing your corpus, and we’ll be looking at some of these throughout the video. In this video I’ll be drawing on some of my own work in student writing across different disciplines to illustrate corpus linguistic techniques.

So, I’m going to launch straight into an example showing the value of quite a small corpus of a few hundred thousand words. So here you can see on the right a student assignment in engineering. You can see there are lots of figures and some text, and that’s what you’d see if you were to read it page by page. But when you put it into the corpus, of course it’s all kind of crunched down and you’re just seeing all the words together. The diagrams are tagged, that is to say there’s a place marker put in, and taken out. And in my study I looked at similarities and differences between students with different first languages and also across discipline areas, so the little bit I’m going to look at today focuses on just a small portion of the whole, so a few hundred thousand words in engineering and business.

First of all, here are some statistics from a Chinese undergraduate student in business. So, their first language was Chinese and they’re studying in English in a UK university. What can this tell you? First of all, it tells you the tokens or number of words. And if you look at the top you can see overall, so that’s in the whole corpus. Then you can see in text one, text two, text three, etc., so you can see that the number of tokens, we’ve got 33,000 words overall, so quite small.

The number of types relates to the different varieties of words, so a sentence like the cat sat on the mat is six tokens, six individual words, but just five types because the definite article, the, is repeated. So, type token ratio is a way in to looking at the variety of the lexis used. Does one corpus contain a wider variety of lexis than another? And mean word or sentence length is a way in to looking at the complexity of language. Does one corpus use longer words or longer sentences or paragraphs than another? And finally, you can see the averages per text here, the number of one-letter, two-letter, three-letter words.

Another starting point might be a word list. This is a straightforward list of words in order of frequency in a corpus. So, this was a task in the worksheet before this video. If you didn’t do it then, maybe have a think now. What are the most common ten words in written English taken from web sources? And here they are, actually top 20. So, you’ve got common grammatical words, the, be, to, of, and, that occur widely across a variety of texts. So, I compared this with my Chinese student corpus, the same Chinese undergraduate student of business, and you can see straight away there are similar words, so the features in both, and, to, of, a, in, all the little grammatical words, but we also have the hash symbol. And in the software I used here, WordSmith Tools, the hash symbol denotes a number. And we’ve also got the first person I.

This slide shows a plot dispersion also generated by WordSmith Tools, in which we can see the use of the first person singular, or I, across 12 texts. So, going from left to right you can see the number of the text, the filename, the number of words in the text, the number of hits or occurrences of I, the same information per 1000 words, and finally the dispersion. So, from left to right, each little red line shows from the beginning on the left to the end on the right of each file. And if you look at the right-hand side, you’ll see that a number of the texts have a little cluster of red lines at the end, denoting quite a few occurrences of I at the end of the text.

So, why might this be the case? And this occurs more broadly than just in these Chinese undergraduate students in engineer; it occurs in writing by people with English as their first language and across a number of disciplines. Think about what students are asked to do at the end of their writing. And to answer this question, you might need to drill down and read some of the texts to find out what’s going on. So here it brings in the discourse analysis. And here’s the answer; it’s reflective writing. So, frequently, at the end of an undergraduate’s assignment, students are asked to reflect on the process of carrying out the work or of writing the assignment or of working in a team, hence a lot of occurrences of I.

[0:06:42]

Using the same dataset of student writing, I looked at keywords. So, these are defined as occurring more often than would be expected by chance in comparison with a reference corpus. And here are keywords in Chinese engineering writing when compared to students with a first language of English, also in engineering. Most are maybe not all that interesting because they simply indicate the topics that several students chose to cover in their assignments, but here’s that hash sign again. So, are Chinese engineering students using more numbers than L1 English students and why? And this seems worth investigating further. When I looked into it and looked at the assignments in more detail, I saw that the Chinese students tended to use more numbered lists than the L1 English students.

I also looked at key clusters, and in WordSmith Tools, cluster is a term used to refer to lexical chunks or words that occur together. Here I’ve pulled out the four-word chunks. We’ve got the positively key ones in black and then the negatively key ones in red, so the ones that occur more often in the L1 English students’ writing than the L1 Chinese ones.

So, here we get some insight into the discourse markers and the signposting used more frequently by the Chinese students than the British students, so, for example, on the other hand, last but not least, according to the equation. And is it the case that Chinese students use more equations than the British students or is it that they simply use this particular expression to refer to their equations. And the corpus can’t really tell us this; we need to dig a little bit deeper and read the text to find out.

This slide looks at the area of semantic tagging. So, here I used Wmatrix to explore the areas that were key in semantic terms, in meaning, in one corpus compared to another as a way into the analysis. So, for example, the word enormous might not occur often enough to make it to a keyword list, but if you explore the category of large size then enormous combined with big, huge, etc., might make that category of large size key in one corpus compared to another.

Here we’re looking at semantic domains within engineer, so I did a comparison of engineering, all student undergraduate writing in engineering, no matter which first language the students had, compared to all student writing at undergraduate level, and what were the differences? So, if you have a look at these, is there anything unexpected? So, you would expect the aboutness of engineering to including cause and effect, electricity, mathematics, measurement, etc. Are there any perhaps unexpected features?

Here are two areas I noticed from the semantic domains, so anatomy and physiology came up because the word appendix occurred quite often. So, of course, it’s about student writing, it’s not an appendix as part of the body; it’s an appendix to a student assignment, and they apparently occurred more often in the engineering corpus. The word worry was misclassified from the words stress and tension, which are technical terms in engineering. So this, for me, shows that you need to eyeball your output. You need to explore the concordance lines and really see what’s going on. So, I hope this has shown that the semantic domains can tell you something about your corpus, and it provides another way of exploring it, but you do need to kind of delve a bit deeper into what’s going on, look a collocate lists, concordance lists, maybe read a few texts to check that what you think is going on is in fact the case.

And we could do a lot more here. We could look at how change is talked about, how numbers are used in the corpus, how cause and effect language is maybe different from other disciplines. And I think here it’s helpful to use several corpus linguistic techniques combined with close reading to see what’s going on in your corpus and try to triangulate, find the same thing out through each different window or way of examining your corpus. And if you have different findings, well, that’s interesting in itself. Has something gone wrong? Do you need to look at things more closely? So, looking at new avenues into your dataset.

National Centre for Research Methods (NCRM)  
Social Sciences  
Murray Building (Bldg 58)  
University of Southampton  
Southampton SO17 1BJ  
United Kingdom

**Web** www.ncrm.ac.uk   
**Email** info@ncrm.ac.uk  
**Tel** +44 23 8059 4539  
**Twitter** @NCRMUK