Lesson 3

Wordlists and keyword lists

There are many ways in which a corpus can be examined and two of the most frequently used tools in the text analysis software tool boxes are wordlists and keyword lists.

Wordlists

- In his consideration of the effects of corpus-based methods on language study, Mike Scott mentions two causes of what he calls an 'upheaval' (Scott and Tribble, 2006: 5): the first are the technological innovations that permit us to plough through vast quantities of text in a short time and to reduce it or 'boil it down' to lists and concordance lines; the second, meanwhile, is the way the pattern-perceiving predisposition of the brain comes into play when it examines such lists. To be useful for investigation, corpora the size of SiBol (see below) need a considerable degree of 'boiling down', and, inevitably, different patterns stand out to different researchers.
    - '[W]here one examines the boiled-down extract, the list of words, the concordance. It is here that something not far different from the sometimes-scorned "intuition" comes in. This is imagination. Insight. Human beings are unable to see shapes, lists, displays, or sets without insight, without seeing in them "patterns". It seems to be a characteristic of the homo sapiens mind that it is often unable to see things "as they are" but imposes on them a tendency, a trend, a pattern.' (Scott and Tribble 2006: 6)

> A word list at first sight is a confusing animal, with its high-frequency items rising up like tusks and its hapax legomena lying as flat as fur; its patterns are weird and wonderful. Beneath the surface though its DNA reveals numerous regularities which can be useful to language researchers searching for patterns of importance in their own text corpora. (Scott and Tribble, 2006: 31)

As Scott argues (2006: 25), a wordlist will nearly always contain, firstly, a small number of highly used items, the most frequently used of these being grammatical items, followed by a long list of items which occur very infrequently.

Keywords

- it is possible to compare the frequencies in one wordlist against another in order to determine which words occur statistically more often in wordlist A when compared with wordlist B and vice versa. Then all of the words that occur more often than expected in one file when compared to another are compiled together into another list, called a keyword list.

> It is this keyword list which is likely to be more useful in suggesting lexical items that could warrant further examination. A keyword list therefore gives a measure of saliency, whereas a simple word list only provides frequency.

(Baker, see Word frequency and keyword extraction article)

Example of research: Time lapse CADS (see slides and handouts)

Here is an example of research into newspaper discourse which used keywords as a way in to the large amount of data.

Keywords with Wordsmith: This allows us to compare the *relative* frequency of items in any corpus with reference to another corpus. The analyst first prepares a list of the items in the first corpus, known as the target corpus, in order of their absolute frequency, using the *Wordlist* tool. The same procedure is followed for the second corpus, known as the reference corpus. The *Keywords* tool can then compare the contents of the two lists and those items which occur statistically significantly (using chi-squared or log-likelihood tests) more frequently in the first list are themselves put in an ordered list. The more statistically significant the item, the more *key* it is, the higher it is placed on the list. This keyword list, providing an ordered series of items which are *salient* in one corpus compared to another corpus, is likely to suggest items which warrant further investigation (Baker, 2006: 125). The procedure can then be repeated but by inverting the two corpora to reveal the items which are salient in the second corpus.

Following this methodology, then, three lists of keywords were produced, one of the salient items in 2010 newspapers relative to 1993, one of the 2005 newspapers relative to 1993 and the third of the key items from the 1993 newspapers relative to 2010 and 2005 combined. Even when setting the *WordSmith Keyword* statistical significance setting at the most rigorous level envisaged (that is by setting the *lowest* p-value available, namely p = 10^15), the two corpora being compared were so large that each list contains over 7,000 items. However, for practical purposes the first 2,500 items in each list were taken into consideration.

Modern-diachronic corpus-assisted discourse studies (MD-CADS): using the SiBol sister newspaper corpora

Modern diachronic corpus-assisted discourse studies (MD-CADS) employs large corpora of a parallel structure and content from different moments of contemporary time in order to track changes in modern language usage but also social, cultural and political changes over modern times, as reflected in language.

Using recently compiled much larger corpora from different time periods, each containing over 100 million words of newspaper texts, corpora constructed to be as similar in content, composition and structure as possible, the first set of MD-CADS work from the SiBol group, were able to study a variety of fine grained lexico-grammatical changes, secondly typical discourse practices within a discourse type (that is typical ways of saying things) and also compared earlier with more recent attitudes to certain social, cultural and political phenomenon, as projected by the mainstream UK quality press, (Partington ed. 2010). The corpora, a complete year of Times, Telegraph and Guardian from 1993 and then from 2005 were enhanced by the addition of Port2010 compiled by Charlotte Taylor. The intent has been to track how language patterns and meanings, as well as the discourse practices these might reveal, can change over comparatively brief periods of modern times

Grammatical developments observed include changes in time of an increase in the use of personal pronouns and verb contractions and a decline in the use of honorifics as well as differences in choice of linkers and modals. Other developments observed included changes in evidentiality, that is, how newspapers give evidence for the claims they make, the ways in which science is reported and, above all, that the UK so-called 'quality' papers are adopting many of the language behaviours once thought more typical of their tabloid rivals including an increase in hyperbolic evaluation and vague language.

Sociopolitical and cultural studies also saw changes in what the UK papers considered moral and immoral over the period and what were favoured "moral panics", changes and similarities in the way BOY and GIRL were represented and the appearance of disturbing new representations of antisemitism in Europe and the UK.