AHRC ICT Methods Network Expert Seminar on Linguistics

Word Frequency and Keyword Extraction

Lancaster House Hotel, Lancaster, Thursday September 8th 2005

ABSTRACTS

Word Frequency: Use or misuse?

John M. Kirk, *Queen's University Belfast, Northern Ireland.*

This paper will not be concerned with statistical treatments of word frequency beyond percentage distributions and relativised frequencies per thousands or millions of words. Its primary concern will be with frequency as a property of data, adopting a critical look at statements such as 'each text comprises 2,000 words'. It will be concerned with words as tokens, types and lemmatised types; the range of functions and meanings of words; and words and lexemes. It will consider words of low frequency as well as of high frequency.

In its critical section, it will ask whether word frequencies are self-explanatory or need explanation, and whether approximation is as useful as precision. It refers to a range of well-known corpora of English as well as the three corpora which I have compiled: *Corpus of Dramatic Texts in Scots*, the *Northern Ireland Transcribed Corpus of Speech (NITCS)*, and the Irish component of the *International Corpus of English (ICE-Ireland)*.

Word Frequency, Statistical Stylistics, and Authorship Attribution.

David Hoover, *New York University, USA.*

The availability of large corpora and of electronic texts has renewed interest in the venerable topic of word frequency. Innovations in analytic techniques and in the ways word frequencies are selected for analysis have also been instrumental. It is these upon which I will focus here. Authorship attribution and statistical stylistics have, until recently, typically been based upon fewer than the 100 most frequent words of a corpus. These words – almost exclusively function words – are attractive because they are so frequent that they account for most of the running words of a text, and because such words have been assumed to be especially resistant to intentional manipulation by an author, so that their frequencies should reveal authorial habits that remain relatively constant across a variety of texts.

Recent work on style variation has suggested that selecting words because of their frequency in sections of texts rather than in the entire corpus is more effective in capturing stylistic shifts. Removing words that are frequent overall only because they are very frequent in a single text has also been shown to dramatically improve the accuracy of an analysis. Another trend has been to increase the number of words analyzed to as many as the 6000 most frequent words, a point at which almost all the words of the text are included and almost all are content words. Finally, following the innovative work of John Burrows, there is a great deal of current interest in Delta, a new measure of the differences between texts that is based upon comparing how different the texts are from the mean for the entire corpus. Further refinements in the selection of words for analysis and in alternative formulas for calculating Delta suggest that further improvements in the accuracy may be possible and that we may be nearing a theoretical explanation of how and why word frequency analysis is able to capture authorship and style.

I will be discussing these issues mainly with reference to a 2,000,000 word corpus of contemporary American poetry and a much larger corpus of 46 Victorian novels.

Word frequency in Context: Alternative architectures to examine related words, register variation, and historical change.

Mark Davies, *Brigham Young University, USA.*

The use of relational databases that are composed of the frequency of n-gram in a given corpus allows users to quickly and easily examine word frequency. Perhaps the first large corpus to use such an approach was the 100 million word *Corpus del Español*, which was created in 2002 (www.corpusdelespanol.org/). This was followed by two BNC-based 100 million word corpora that were modelled on the same architecture: *Phrases in English* (pie.usna.edu) and *Variation in English Words and Phrases* (*VIEW*; view.byu.edu), as well as a 40 million word *Corpus of Historical English* (view.byu.edu/che)

The relational database/n-grams architecture allows simple word frequency queries such as the following (all of which can be carried out on a 100 million word corpus in 1-2 seconds):

- Overall frequency of a given word, set of words, phrase, or substring in the corpus
- "Slot-based" queries; e.g. the most common nouns one "slot" after *mysterious*, or z-score ranked words immediately preceding *chair*
- Wide-range collocates; e.g. the most common nouns within a ten word window (left or right) of *string* or *broken*

In addition, however, the architecture that we have used for *VIEW* and the *Corpus of Historical English* allows several other types of queries that cannot be carried out directly with competing architectures (e.g. *SARA/XARA*, the *IMS Corpus Workbench*, or the *Phrases in English* architecture), including the following:

- Comparison of frequency with related words; e.g. nouns occurring immediately after *utter* but not after *complete* or *sheer*, or adjectives within ten words of *woman* but not *man*
- One simple query to find the frequency of words in separate databases, such as user-defined, customized lists (clothing, emotions, technology terms, etc) or synsets from WordNet
- Register variation; e.g. all verbs or all words ending in *ble or all three-word lexical bundles that are more common in academic texts than in fiction, or in legal or medical texts
- Historical variation; e.g. words, phrases, or collocates of a given word or part of speech, which are more common in the 1900s than in the 1800s

Finally, even within the relational database/n-grams approach, there are competing architectures that favour certain types of queries over others, and we will briefly consider some of these issues.

Issues for Historical Corpora: First catch your word.

Christian Kay, *University of Glasgow, Scotland.*

The *Historical Thesaurus of English* (*THE;* www.arts.gla.ac.uk/sesll/englang/thesaur/homepage.htm) is a semantic index to the *Oxford English Dictionary* (*OED* 1884-; *OED Online* 2000-) supplemented by Old English materials published separately in *A Thesaurus of Old English* (Roberts, Kay, Grundy, 2000). Word senses are organised in a hierarchy of categories and subcategories, with up to fourteen levels of delicacy. The material is held in a database and first steps towards Internet publication are being taken by an AHRC-ICT Strategy Project creating searches for use in a range of humanities disciplines (Smith, Horobin, Kay, n.d.). The main problem which besets searching historical texts is that of variable spelling – the further one goes back in time, the worse it gets.

Similar problems affect texts in non-standard varieties, as experience of the *Scottish Corpus of Texts and Speech* (*SCOTS;* www.scottishcorpus.ac.uk/) and the *Dictionary of the Scots Language* (DSL; www.dsl.ac.uk/dsl/) demonstrate. Dictionary headwords lemmatize common variants but are by no means comprehensive; an alternative may be a rule-based system which predicts possibilities. Corpora have further problems in that lemmatization may not solve problems of homonymy and polysemy. The paper will suggest ways of addressing these problems using the resources described above.

References
The Oxford English Dictionary, 1884-1933 ed. by Murray, Sir James A. H.; Bradley, Henry; Craigie, Sir William A. and Onions, Charles T.; *Supplement*, 1972-1986 ed. by Burchfield, Robert W., 2nd edn, 1989, ed. by Simpson, John A. and Weiner, Edmund S. C.; *Additions Series*, 1993-1997, ed. by Simpson, John A.; Weiner, Edmund S. C. and Proffitt, Michael; 3rd edn (in progress) *OED Online*, March 2000-, ed. by Simpson, John A. (Oxford: Oxford University Press).

Roberts, Jane and Kay, Christian with Grundy, Lynne, *A Thesaurus of Old English*, King's College London Medieval Studies XI, 2 vols (London, 1995), Second impression (Amsterdam: Rodopi, 2000). An electronic version, supported by British Academy LRG-37362, may be seen at www.arts.gla.ac.uk/sesll/englang/thesaur/toe1.htm.

Smith, Jeremy; Horobin, Simon; and Kay, Christian, *Lexical Searches for the Arts and Humanities*, AR112456.

In Search of a Bad Reference Corpus

Mike Scott, *University of Liverpool, UK.*

What are the tolerable limits of similarity between a reference corpus and a node text for the generation of a useful set of keywords? There is of course considerable subjectivity in the notion of usefulness, which will vary according to research goals which cannot in general be predicted with certainty. Nevertheless, the aim here is to explore the ways in which the similarity between reference corpus and node text vary on various important dimensions, such as size in tokens, similarity of text-type, similarity of historical period, similarity of subject-matter.

This presentation starts from the formula proposed by Berber Sardinha (2004, 101-3) which suggests that the larger the reference corpus, the more keywords will be detected, and his formula for predicting the number of keywords produced with a given text and reference corpus. It also considers his recommendation that a reference corpus should be about five times the size of the node text.

Using a series of reference corpora, the paper compares keyword results in relation to specific texts. The aim is to identify not, as one might imagine, the characteristics of the good reference corpus, but the limits defining a poor one, since in many cases, e.g. the analysis of a dead language or a restricted corpus, the chance of accessing a good reference corpus is slim.

References
Berber Sardinha, A. P., *Lingüística de Corpus* (São Paulo, Brazil: Manole, 2004).

Keywords and Moral Panics: Mary Whitehouse and media censorship

Tony McEnery, *Lancaster University, UK.*

In this paper I will use an analytical framework based around the use of keywords to investigate the moral panic encoded in the writings of Mary Whitehouse in the 1960s and 70s in Britain. In doing so, I will be using keywords as a way of focusing on the *aboutness* of the moral panic, and a study of patterns of colligation and collocation to explore convergence in these texts.

Subsequently, I will consider the issue of bad language and consider how bad language was represented by Whitehouse's organisation VALA (Viewers and Listeners' Association). The paper will consider throughout how the moral panic in the corpus of Whitehouse's writings compares to that in the writings of the Societies for the Reformation of Manners, religious organisations in the seventeenth century which opposed bad language (among other behaviours). The point of departure for all aspects of this investigation is the question of moral panics and the use of keywords to explore them.

‘The question is, how cruel is it?’ Keywords in debates on fox hunting in the British House of Commons.

Paul Baker, *Lancaster University, UK.*

A small corpus of 130,000 words consisting of debates on fox hunting which took place in the British House of Commons in 2002 and 2003 was collected and then subjected to a keywords analysis. The corpus was split into two sub-corpora depending on whether speakers argued for or against fox hunting to be banned. The sub-corpora were compared together, resulting in separate keyword lists for each. Proper nouns and words relating to the debate's context (parliament) were removed from the lists prior to analysis.

This paper examines a number of keywords in detail, using concordance analyses, in order to identify different discourses (ways of looking at the world) that speakers access in order to persuade others of their point of view.

I also explore additional ways of using keyness to find salient language differences in texts, for example, by looking at key clusters and key semantic categories as well as comparing the whole corpus to a reference corpus of general British English.

Love – a familiar or a devil? An exploration of key domains in Shakespeare's Comedies and Tragedies

Dawn Archer, Jonathan Culpeper, Paul Rayson, *Universities of Central Lancashire and Lancaster, UK.*

Love is a common theme in Shakespeare’s works. In this paper, we show how the UCREL Semantic Annotation Scheme (henceforth USAS), a software program for automatic dictionary-based content analysis, can help us to explore the semantic field of ‘love’ within a selection of Shakespeare’s plays. Specifically, we will explore three love-tragedies (*Othello*, *Antony and Cleopatra*, and *Romeo and Juliet*) and three love-comedies (*A Midsummer Night’s Dream*, *The Two Gentlemen of Verona* and *As You Like It*) to determine differences in their (re)presentation of ‘love’. We will also discuss how the semantic field of ‘love’ co-occurs with different domains in the plays, and assess the implications this has on our understanding of ‘love’ as a concept.

This research builds on (i) Jonathan Culpeper’s work on keywords in Shakespeare, using Wordsmith (Culpeper 2002), (ii) Paul Rayson’s comparisons of key word and key domain analysis (Rayson 2003), and (iii) Dawn Archer and Paul Rayson’s work on the identification of key domains in refugee literature, using USAS (Archer and Rayson 2004).

References

Archer, D. and Rayson, P., ‘Using the UCREL automated semantic analysis system to investigate differing concerns in refugee literature’, in Deegan, M.; Hunyadi, L. and Short, H. (eds.) *The Keyword Project: Unlocking Content Through Computational Linguistics* (Office for Humanities Communication Publications, 18, London, forthcoming).

Culpeper, J., 'Computers, language and characterisation: An analysis of six characters in *Romeo and Juliet*', in Marttala, Ulla Melander; Ostman, Carin and Kyto, Merja (eds), *Papers from the ASLA Symposium: Conversation in Life and Literature* (Association Suedoise de Linguistique Appliquee, 15, Uppsala, 2002) pp. 11-30.

Rayson, P. ‘Matrix: A Statistical Method and Software Tool for Linguistic Analysis Through Corpus Comparison’ Ph.D. thesis (Lancaster University, 2003).