AHRC ICT Methods Network Expert Seminar on Linguistics

WORD FREQUENCY AND KEYWORD EXTRACTION

Hosted by Professor Tony McEnery, Department of Linguistics and English Language, Lancaster University.

Rapporteur's Report (draft)

Marilyn Deegan

The introduction to the seminar by Professor Tony McEnery situated the importance of applied linguistics in general and contrasting word frequency information in particular within the arts and humanities, stressing the important roles that both play. He flagged some of the key issues: historians approaching primary source documents are likely to benefit from knowledge about discourse analysis, for instance, and this asks the wider question of how we approach documents and language. The current availability of large quantities of information in digital form, too, means that it may be important to be able to gist ever larger collections of data, something of importance in most academic fields, as well as in many applications outside the academic world: business, defence, any other arenas where it is vital that the key information in large collections is grasped rapidly. He also pointed out the value of corpus linguistics in media representations of issues and groups in society. The papers to be presented at the seminar also link out to other areas of study: literature, politics, media studies, among others.

The world is awash with text, much of it available digitally, and making some sense of this plethora needs some structured algorithmic approaches. There are many new corpora which have not been built or gathered according to the principles of corpus linguistics, and which are therefore more fuzzy, but which are nevertheless interesting and important. Linguistic and statistical techniques can be used to aggregate information that is hidden in the mass. Sizes of corpora are now vastly larger than those created by corpus linguistic principles: the *British National Corpus*, completed in 1994, contains more than 100 million words, while the *Google Print* initiative which plans to digitise seven million books will yield 50 billion words. Given the size of some of these corpora, McEnery emphasised the importance of context in the interpretation of results from word frequency investigations. He also discussed the refinement of tools used in frequency studies so that they could be used both to give overall pictures of key concepts in texts, as well as allowing close-up views. Keyword and frequency studies therefore have very broad applicability both within and outside the academic world.

The primary concern of John Kirk's paper 'Word Frequency: Use or misuse?' was with frequency as a property of data, and he offered a critical analysis of statements such as 'each text comprises 2,000 words'. The presentation was largely concerned with words as tokens, types and lemmatised types; the range of functions and meanings of words; and words and lexemes. Kirk questioned whether word frequencies were self-explanatory or in need of further explanation, and whether approximation could be as useful as precision. He referred to a range of well-known corpora of English as well as three corpora which he had compiled. He also discussed the modern version of authorship studies: its use in the detection of plagiarism, and he discussed the contribution made to linguistic theory by word frequency studies.

David Hoover, in ‘Word Frequency, Statistical Stylistics, and Authorship Attribution’, suggested that the availability of large corpora and of electronic texts has renewed interest in the topic of word frequency, and pointed out that innovations in analytic techniques and in the ways word frequencies are selected for analysis have also been instrumental in this revival. 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 seen as 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, however, 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 analysed 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. Hoover also discussed the innovative work of John Burrows in producing his *Delta* technique, a new measure of the differences between texts 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.

Hoover discussed these issues with reference to a 2,000,000 word corpus of contemporary American poetry and a much larger corpus of 46 Victorian novels.

Mark Davies in ‘Word Frequency in Context: Alternative architectures to examine related words, register variation, and historical change’ discussed some alternatives to techniques based on word searching. He proposed that architectures based on relational databases and n-gram frequencies dramatically improved performance in the searching of corpora, and suggested that simple word frequency queries can be carried out on a 100 million word corpus in 1-2 seconds. He described a number of corpora that have been created using this approach, including the 100 million word *Corpus del Español*, which was created in 2002, and two BNC-based 100 million word corpora that were modelled on the same architecture: *Phrases in English and Variation in English Words and Phrases* (*VIEW*), as well as the 40 million word *Corpus of Historical English*.

He also discussed the fact that, even within the relational database/n-grams approach, there are competing architectures that favour certain types of queries over others.

In contrast to the other speakers, Christian Kay in ‘Issues for Historical Corpora: First catch your word’, was discussing historical rather than modern corpora, with particular reference to the *Historical Thesaurus of English* and *A Thesaurus of Old English*. The main problem which besets searching historical texts, according to Kay, is that of variable spelling – the further one goes back in time, the worse it gets. This is also a critical issue in texts in non-standard varieties, as experience of the *Scottish Corpus of Texts and Speech* (*SCOTS*) and the *Dictionary of the Scots Language* (*DSL*) demonstrate. Kay pointed out that historians are also having to face these problems in texts, and she also discussed the relationship between e-texts (of which there are many) and structured corpora (of which there are few). The issues raised in this paper are critical for keyword extraction and word frequency work, and thesauri such as those discussed by Kay which give variant spelling forms could be of enormous benefit in building access tools for problematic corpora.

The topic tackled by Mike Scott was in the areas of reference corpora ('In Search of a Bad Reference Corpus'). Scott set out to explore the tolerable limits of similarity between a reference corpus and a node text for the generation of a useful set of keywords. As he suggested, there is considerable subjectivity in the notion of usefulness, which will vary according to research goals which cannot in general be predicted with certainty. His expressed aim was to explore the ways in which the similarity between a reference corpus and a 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 began with the formula proposed by Berber Sardinha 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 considered 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 compared keyword results in relation to specific texts. The aim was 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. Surprisingly, even relatively restricted reference corpora can give good results in keyword extraction, and Scott concluded that a small reference corpus containing a mixture of texts performed better than larger corpora with more homogenous texts.

Tony McEnery's paper, 'Keywords and Moral Panics: Mary Whitehouse and media censorship', proposed 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. McEnery used keywords as a way of focusing on the aboutness of the moral panic, and looked at patterns of colligation and collocation to explore convergence in these texts. He then considered the issue of bad language and looked at how bad language was represented by Whitehouse's organisation VALA (Viewers and Listeners' Association). The paper also examined 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 was the question of moral panics and the use of keywords to explore them. McEnery explored too a related topic, that of 'key' keywords, a distillation which could give more information on the aboutness of texts, and discussed the degree to which analysis informs the way the research is modelled, rather than a pre-determined model dictating the analysis.

Paul Baker in 'The question is, how cruel is it?' looked at keywords in debates on fox hunting in the British House of Commons. Baker created 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. This was then subjected to a keyword 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. The research questions Baker explored were: How is language used in the debate to construct different discourses about fox-hunting? What rhetorical strategies are used in the debate? There were some interesting and surprising findings about the differences between the pro- and anti-hunt lobby, and Baker concluded that keywords offered a potentially useful way of focusing researcher attention on aspects of a text or corpus, but that care should be taken not to over-focus on difference/presence at the expense of similarity/absence. Multiple reference corpora need to be used to gain the fullest possible picture.

Dawn Archer, Jonathan Culpeper, and Paul Rayson explored some key domains in Shakespeare's Comedies and Tragedies in ‘Love – a Familiar or a Devil?’ Love is a common theme in Shakespeare’s works, and the presenters showed how the *UCREL Semantic Annotation Scheme* (henceforth *USAS*), a software program for automatic dictionary-based content analysis, helped them to explore the semantic field of ‘love’ within a selection of Shakespeare’s plays. Specifically, they explored 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’. They also discussed how the semantic field of ‘love’ co-occurs with different domains in the plays, and assessed the implications this has on the understanding of ‘love’ as a concept in Shakespeare.

Their key findings were as follows. First of all, there is a marked difference in the occurrence of the concept ‘love’ between comedies and tragedies, it being underused in the tragedies, which focus more on war, death, and other related matters. Where it is used in the tragedies, the representation is much darker. There was also found to be some degree of gender bias. They concluded that the analysis of key domains could provide some useful results, enabling links across different semantic fields to be spotted. Moreover, the findings allowed the team to see where the tools they were using could be usefully refined.

The general discussions at the end of the Seminar ranged widely across the issues that had been presented during the day. There was some debate about the overlap between corpus linguistics and information retrieval, and Tony McEnery described ‘keyness’ as a simple and robust model of contrasting word frequency lists. This led to a debate about the consistency with which keyword extraction could be applied consistently: what are the cut-off points in selecting just what *is* a keyword? It was pointed out that it is vital to apply keyword selection, and the importance of context comes in here. Objectivity is another critical issue: machine-generation of keywords is interesting, but the results need interpretation, which is where notions of bias and objectivity come into play. Any selection or deselection of keywords is antithetical to the supposed objectivity of the machine, but there is no such thing as bias-free research or intuition-free linguistics.

Marilyn Deegan concluded the workshop with a summary of some of the key questions that had been debated and with some issues that had not been debated. First of all, she pointed out the degree of cohesion that there had been between many of the researchers in terms of theoretic underpinnings and findings, despite the use of very diverse corpora is terms of size, genre, period, degree of design versus randomness. She also suggested that there should be some work done to draw out what advanced methods were being used by this community, in order that they could be promulgated to/used by a wider community. There had been some discussion about the wide applicability of the methods, but it would also be useful to in initiate some discussions about the composition and range of that wider community.

Deegan also speculated upon what influence e-Science and the Grid might have in corpus linguistics and keyword extraction: could ever-larger corpora be analysed? Would it be possible to analyse distributed corpora? Could tools be made interoperable over Grid networks? She finally pointed out the work that is being carried out in the commercial world. The mining of large volumes of unstructured information is a key commercial research area, given the amount of textual information currently available. IBM’s *Unstructured Information Management Architecture*, for instance, uses combinations of semantic analysis and search components to find information in unstructured texts. Other companies such as nstein offer ‘document intelligence’ in the areas of e-publishing, homeland security, and the corporate world. Keyword extraction is a big business as well as a vitally important academic research area.