TITLE: From key words to key semantic domains

AUTHOR: Paul Rayson

AFFILIATION: Lancaster University

CONTACT DETAILS:

Computing Department,

Infolab21,

Lancaster University,

Lancaster,

LA1 4WA.

Tel: +44 1524 510357

Fax: +44 1524 510492

Email: paul@comp.lancs.ac.uk

REVISED VERSION 29th August 2008

NUMBER OF WORDS : 10,159

TITLE: From key words to key semantic domains

KEY WORDS: key words, POS tagging, semantic annotation, data-driven

ABSTRACT

This paper reports the extension of the key words method for the comparison of corpora. Using automatic tagging software that assigns part-of-speech and semantic field (domain) tags, a method is described which permits the extraction of key domains by applying the keyness calculation to tag frequency lists. The combination of the key words and key domains methods is shown to allow macroscopic analysis (the study of the characteristics of whole texts or varieties of language) to inform the microscopic level (focussing on the use of a particular linguistic feature) and thereby suggesting those linguistic features which should be investigated further. The resulting *data-driven* approach presented here combines elements of both the *corpus-based* and *corpus-driven* paradigms in corpus linguistics. A web-based tool, Wmatrix, implementing the proposed method is applied in a case study: the comparison of UK 2001 general election manifestos of the Labour and Liberal Democratic parties.

1. Introduction

The methodology used by corpus linguistics researchers typically proceeds along the following lines: it begins with the identification of a research question, continues with

building and annotating a corpus with which to investigate the topic, and finishes with the retrieval, extraction and interpretation of information from the corpus which may help the researcher to answer the research question or confirm the parameters of the model. In some cases, the process may be an iterative one, where, following the interpretation of the results, some refinement is needed on the research question or annotation of the corpus. This process model is set out in five core steps below:

1. *Question*: Devise a research question or model
2. *Build*: Corpus design and compilation
3. *Annotate*: Manual or automatic analysis of the corpus
4. *Retrieve*: Quantitative and qualitative analyses of the corpus
5. *Interpret*: Manual interpretation of the results or confirmation of the accuracy of the model

There are at least three further stages to the research process that are typical across many, if not all, disciplines, although we will not consider these other than to list them here for completeness:

6. *Output*: Distil the research into a paper or presentation
7. *Disseminate*: Pass the paper to a publisher for printing, or submit to a conference for presentation
8. *Feedback*: Reviews of papers or presentations and citation practice influence the direction of future research

There are two general kinds of research question (Step 1 above) that can be investigated in linguistics using a corpus-based paradigm. First, research can focus on the use of a particular linguistic feature, possibly a word, lemma, multiword expression or a grammatical construction. I will call this Type I. Secondly, one can examine the characteristics of whole texts or varieties of language, and I will call this Type II. These two types are sometimes referred to as microscopic (Type I) and macroscopic (Type II), for example see Biber (1988: 61). Traditionally, studies tend to focus on Type I and examine linguistic (lexical or grammatical associations of the feature), and non-linguistic aspects (distribution of the feature across different types of texts or speech). Type II inverts this relationship by investigating, for example, register variation across text, by examining how certain features or groups of features characterise a text.

Increasingly, researchers no longer have to build and annotate their own corpus material (Steps 2 and 3 above), although this is usually the case with problem-oriented tagging, for example, error tagging of learner data (Granger 1999). Instead they can use precompiled and annotated corpora that are available 'off-the-shelf' (Meyer 1991). To retrieve and interpret data from corpora (Steps 4 and 5), there are two well-known methods used in corpus linguistics. These are frequency profiling and concordancing.

This process model, as described above, is in line with Leech's (1992) view of the corpus linguistics paradigm. Leech argues that the corpus-based methodology conforms to standards commonly ascribed to 'the scientific method': falsifiability, completeness, simplicity, strength, and objectivity.

There are many examples of both types of research question in many conference publications, journal papers and edited collections that have appeared. Common to both is the prior selection of which linguistic features to study. The method proposed in this paper allows a different approach: decisions on which linguistic features are important or should be studied further are made on the basis of information extracted from the data itself; in other words, it is *data-driven*. I will call this Type III. It combines the approaches of Types I and II by first focussing on whole texts and then suggesting specific linguistic features to study in further detail. In other words, the ordering of the five main steps above will change to the following (with iteration back from Step 4 to Step 3, which enables refinement of the research question following a retrieval step):

1. *Build*: Corpus design and compilation
2. *Annotate*: Manual or automatic analysis of the corpus
3. *Retrieve*: Quantitative and qualitative analyses of the corpus
4. *Question*: Devise a research question or model (iteration back to Step 3)
5. *Interpret*: Manual interpretation of the results or confirmation of the accuracy of the model

My Type III process model shown here is similar to that of *corpus-driven* linguistics as presented by Tognini-Bonelli (2001: 85), in which the corpus is the main informant (Francis, 1993). However, I decided to use the term *data-driven* to distinguish my approach from that of *corpus-driven* linguistics, presented by Tognini-Bonelli (2001: xi). The corpus-driven approach questions the "underlying assumptions behind many well established theoretical positions" (2001: 48) stating that they need to be re-established

or replaced based on evidence from corpora. In the corpus-driven approach, Stubbs (1993: 17) notes that the traditional POS system "is under attack". In this study, I will rely on pre-existing part-of-speech (POS) tagsets for example. Hence the Type III, data-driven approach, presented here combines elements of both the corpus-based and corpus-driven paradigms.

This paper is structured as follows. Section 2 of this paper places the work in context of previous research into systematic approaches to corpus comparison. Section 3 introduces the details of the method itself and a case study is presented in Section 4. Finally, I reflect on possible drawbacks and future work in the conclusion, Section 5.

2. Related work

There are some existing examples of Type III studies. Here I will examine three: Ringbom (1998), Hoffmann and Lehmann (2000), and Leech and Fallon (1992).

Ringbom (1998) investigated advanced-learner language in the International Corpus of Learner English (ICLE) by comparing the essays produced by learners to those of native speakers in terms of word frequencies. There were certain problems with this approach as identified by Ringbom. First, there was the assumption that the writing of American and British students form a reasonable norm of argumentative essay writing. Then there was the problem of the ICLE subcorpora being relatively small (roughly 100,000 words each). Ringbom thus restricted the study to high frequency items and reasoned that "if there are fewer than 20 actual occurrences of a word or phrase in such small corpora,

not much can be generalised about the writer's use of this aspect of language". I have identified this as Type III since Ringbom selected two verbs (*get* and *think*) for further study based on their overuse in frequency terms in the non-native speaker corpora when compared to the native speaker data.

I have classified Hoffmann and Lehmann (2000) as Type III since they used collocation evidence from the British National Corpus to select pairs of related words that were then used in a study to discover native and non-native speakers' familiarity with the word pairs. However, they did not pursue the usual Type I path of performing a more in-depth linguistic analysis on the collocates that they discovered. Instead, the paper focuses on analysing the results of the familiarity questionnaire. Due to the large size of the corpus, they selected collocation pairs with less than 100 occurrences to avoid problems of excessive computation. They used the log-likelihood statistic to select 150 collocations.

Leech and Fallon (1992) also describe a two-stage research process which I would categorise as Type III in their examination of cultural differences using corpora of British and American English. Stage one is to use a comparative alphabetical list of word frequencies in the two corpora to select groups of words for further study. This stage examined the Hofland and Johansson (1982) lists of word frequencies in British and American English to select the items marked with significant differences. Stage two made use of a concordance tool to examine the contexts of the selected words from the Brown and LOB corpora. Leech and Fallon cite two main reasons for consulting the concordance lines:

1. To check whether the frequency of the graphic form actually reflected the sense of the word they were interested in.
2. To check that the high frequency of an item was not due to any obvious skewing of its distribution in the corpus.

They describe stage two as requiring "an enormous amount of human labour, and in practice the task had to be simplified". The same issues are faced by other corpus researchers in their studies. The most used current techniques to reduce the number of concordance lines for inspection are that of random sampling, and collocation statistics (arising out of the needs of lexicographers, see Kilgarriff and Tugwell 2002). Leech and Fallon's approach was extended by Oakes and Farrow (2007) to include samples of written English from five additional countries.

What the three examples of Type III studies reviewed here have in common is the use of corpus-based comparative frequency evidence to drive the selection of words for further study. The focus of this paper, then, is on systematic approaches to the comparison of corpora. The key word approach taken by Scott is the most widely cited approach and it is implemented in the Keyword module of WordSmith Tools (Scott 1997, 2001a). Tribble (2000: 79-80) describes the way that WordSmith finds key words as follows:

1. Frequency sorted wordlists are generated for a 'reference' corpus (a collection that is larger than the individual text or collection of texts which will be studied), and for the research text or texts.

2. Each word in the research text is compared with its equivalent in the reference text and the program makes a judgement as to whether or not there is a statistically significant difference between the frequencies of the word in the different corpora. The statistical test evaluates the difference between counts per type and total words in each text and can be based either on a chi-squared test for outstandingness or on a log-likelihood procedure.
3. The wordlist for the research corpus is reordered in terms of the 'keyness' of each word.

Scott (1997) sets a minimum threshold of two occurrences for each word in the research text, although this does result in manually identified key words being omitted from the key words database (Scott, 2001b: 118). The resulting key word list contains two types of key word: *positive* (those which are unusually frequent in the target corpus in comparison with the reference corpus), and *negative* (those which are unusually infrequent in the target corpus). These correspond to the terms *overuse* and *underuse*, used, for example, in the learner corpus literature, e.g. (Granger and Rayson, 1998). Tribble compares the list of positive and negative key words against the frequency list for his corpus and demonstrates the improved usefulness of the key word technique over simple frequencies for extracting interesting lexical items for stylistic studies. Scott also uses the notion of key-key words. These are words that are key in all, or a large percentage, of the texts that are contained in the corpus under investigation. Tribble uses this feature to select lexical items to give pedagogical insights in the study of a particular genre.

Scott (1997) relates his work to that of Raymond Williams in the 1970s in terms of its purpose, but not in terms of its procedure. Williams (1983: 14) selected key words subjectively due to their use in general discussion in 'interesting or difficult ways'. Scott's motivation for his work is a text-focused one, not one aiming to 'characterise a language or a genre, but a language event', and to reveal patterns which construct texts. He argues for the study of texts in their original context with as much detail as possible recorded about the writers or speakers that produced the data. However, he is realistic about recording information about the original circumstances of the language event, such as the mood of the speaker or writer, which may be difficult to recover even for those involved in producing the language. There is no claim that the key words would match those selected by human readers of a text (Scott, 2000a), who may specify a word not even in the text. Scott (2000b) defines an *association* relationship between words, as the co-keyness of both words within the same text, as an alternative to the standard calculation of collocation, which is based on how frequently words occur near to each other. He uses association across a large corpus to investigate the 'aboutness' or content of texts. Tribble (2001) notes that key words regularly occupy potential theme positions in sentences and paragraphs. Scott usually focuses on key open-class words, although his technique may extract closed-class words as well (Scott, 2001b: 126).

Some researchers use the concept of key words in a different way: their key words are not identified statistically. For example, Stubbs (1996: 172) describes *cultural key words*, that is, "words which capture important social and political facts about a community" (Hunston, 2002: 117). The important feature for Stubbs is that these words occur in characteristic collocations, which show the associations and connotations they

have. Stubbs (1996: 166) traces his efforts back to that of Firth (1935) on "focal and pivotal words" and to Williams' book on key words. Stubbs writes that identification of key words will always involve intuition. In his study of language of Euro scepticism in Britain, Teubert (2001) manually selects key words from a pilot corpus and supplements them with significantly frequent collocates of the key words, in a larger corpus. Similarly, Ooi (2000) selects 10 lexical items for their supposed cultural distinctiveness and examines their collocates. The work of Wierzbicka (1997: 16) is also focussed on key words but has no "objective discovery procedure" for them. However, frequency information does play a part in the discovery procedure via checking whether a candidate key word is a common word.

Pre-dating the work of Scott is that of Lyne (1985: 164) who calculates a 'registral value' for each word (instead of using a goodness-of-fit statistic) and sorts on the value to compare frequency data in two corpora. Lyne's goal was to find characteristic vocabulary of French business correspondence. Lyne also proposes a modified registral value which is adjusted for range to filter out technical items.

There are a large number of different measures allowing comparison of frequencies across corpora: Yule's difference coefficient, Pearson's Chi-squared (with various adjustments e.g. Mosteller-Rourke and Yates' continuity correction), Log-likelihood, normalised ratio and Fisher's Exact Test. Further methods allow comparison of ranks of frequency data e.g. the Mann-Whitney test. A survey of such techniques is out of scope for this paper, but the reader is directed to Rayson (2003) for more details (including for

justification of the selection of the Log-likelihood statistic). Oakes (forthcoming) describes other similar metrics originating in Information Retrieval.

Other relevant studies in this area are the works of Biber and Finegan, see for example, Biber (1988), Biber and Finegan (1989), and Biber (1995). These have at their core a comparison of frequency distributions across genres, but use a multi-feature, multi-dimensional methodology, grouping sets of linguistic features associated with a number of factors (called text dimensions). Biber (1988: 63) describes his approach as depending on both the Type I (microscopic) and Type II (macroscopic) research methodologies. He uses the Type II approach to analyse the co-occurrence patterns among the linguistic features, identifying the textual dimensions, and the Type I analyses to interpret these dimensions. The technique proposed by Biber has been widely cited in research articles, but also criticised by Lee (2000) as being linguistically and statistically unsound due to problems to do with the nature of language, the distributional properties of linguistic features and the non-representativeness of corpora. Lee attempted to replicate Biber's dimensions using the same statistical methodology on a four-million-word subset of the BNC, but found that variations in the configuration of the data (relative genre proportions), choice of variables, etc. could distinctly affect the results. This means that Biber's dimensions cannot be considered final. From the point of view of language teachers, Tribble (2000: 78) also points out the practical difficulties in actually using Biber's dimensions or applying them to new texts, due to the necessity of having the research corpus POS-tagged before any analysis can proceed. Tribble then continues his analysis using Scott's key word methodology. The multi-dimensional approach can be seen as inflexible since the features/variables are chosen ahead of the

research question. New features can be chosen, but then one has to repeat the complex analysis procedure, and in so doing may obtain sometimes radically different results, as Lee (2000) demonstrates.

The approach presented in this paper differs from Biber's because it is data-driven: the linguistic features worthy of microscopic analysis are suggested by the macroscopic study, rather than by intuition or previous research studies. My approach mainly aims at the comparison of a small number of text corpora, usually two; one of which may be a normative corpus. Biber's approach considers frequency variation for pre-selected variables across a large number of texts and attempts to situate texts or text genres along several clines of variation. In fact, Xiao and McEnery (2005) demonstrate the similarity of the effect of using the key words procedure to that observed from Biber's multidimensional techniques.

The key words method clearly fits within my Type III category introduced in Section 1. Key words have been used in numerous studies and applied to a wide variety of research questions ranging from sociolinguistics (Baker 2004b) to language education (Scott and Tribble 2006).[1] There has been growing interest shown in key word related events: *AHRC ICT Methods Network Expert Seminar on word frequency and key word extraction*[2] (Lancaster, September 2005) and *Keyness in text*[3] (Siena, June 2007), with the former resulting in a collection edited by Archer (forthcoming). In an extension of the key words approach, Mahlberg (2007) discusses how the keyness technique applied to repeated sequences of words (clusters) can be used to aid the study of literary stylistics.

Berber Sardinha (1999) points out one practical problem with the key words technique: it normally produces more key words than it is possible for the researcher to analyse. He proposes two techniques to reduce the set of words: by selecting a simple majority (i.e. half the number plus one), and by selecting a significant subset (by using the chi-squared test again). Baker (2004a) sounds a note of caution when using the key words technique in relation to three issues:

1. *Difference*: "a key word analysis will focus only on lexical differences, not lexical similarities" (Baker 2004a: 349). Comparing two corpora of gay and lesbian erotic narratives to general corpora such as Frown (Freiburg-Brown) would produce different key words than when comparing the erotic narratives to each other. Hence the choice of reference corpus is important.
2. *Frequency*: a word may be key but only occur in a limited part of a corpus, or relatively low frequency words may be identified as key. Hence, examining the range or dispersion of a key word is recommended.
3. *Sense*: "key words only focus on lexical differences, rather than semantic, grammatical, or functional differences" (Baker 2004a: 354). Baker reports cases where a word is key due to its appearance within a number of distinct meanings, and in addition, where a word is not shown to be key because counting all its meanings together masks the fact one of the meanings is key when counted separately.

The last two points echo the justification made by Leech and Fallon (1992) for thorough examination of concordance lines to investigate for skewed distribution of high frequency words and checking the sense of the word in question.

Gries (2006: 116) also highlights one limitation of corpus variability studies at the lexical level as "they have little or nothing of interest to offer a linguist who is primarily interested in grammatical or other phenomena". Further problems with using frequency lists and key words will be exemplified in the case study described in Section 4.

The method presented in this paper is therefore intended as an extension to the key words procedure rather than a replacement for it. To address the criticisms described here by Berber Sardinha, Gries and Baker, two additional levels of corpus annotation are employed and the keyness method is applied at those levels in addition to the word level. This emerging methodology described in Section 3 also provides a data-driven (Type III) approach to corpus linguistics as introduced in Section 1. The case study presented in Section 4 will exemplify this by highlighting research directions suggested by the method.

3. Extending the keyness method

In this section, I describe the key words method and describe my proposal to apply the same keyness calculation to frequency profiles of higher levels of annotation of the text. Given two corpora that we wish to compare, we produce a set of three frequency lists for each corpus. In the key words approach, this would be only a word frequency list,

but we produce a part-of-speech (POS) and semantic field (domain) frequency list as well. In order to achieve this, each corpus is first tagged using the CLAWS tagger developed at Lancaster (Garside and Smith, 1997). In addition, I use the USAS tagger (Rayson et al 2004b) which automatically assigns semantic fields (domains) to each word or multiword expression in the corpora.[4]

Here I will describe the keyness comparison at the word level. The application of this technique to POS or semantic domain frequency lists is achieved by constructing the contingency table below with tag frequencies rather than word frequencies. As with the key words procedure, due to independence assumptions it is important that the two corpora do not overlap, or that one is not a sub-corpus of the other.

For each entry in the word frequency lists from the two corpora, I calculate the log-likelihood (henceforth LL) statistic. The calculation is performed by constructing a contingency table as in Table 1.

Table 1. Contingency table for log-likelihood calculation

	CORPUS ONE	**CORPUS TWO**	**TOTAL**
Frequency of a word	a	b	a+b
Frequency of other words	c-a	d-b	c+d-a-b
TOTAL	c	d	c+d

Note that the value ‘c’ corresponds to the total number of words in corpus one, and ‘d’ corresponds to the total number of words in corpus two (N values in the formula below). The values ‘a’ and ‘b’ are called the observed values (O). The expected values (E) are calculated according to the following formula:

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i}$$

From the contingency table shown, N1 = c, and N2 = d. So, for this word, E1 = c × (a+b) / (c+d) and E2 = d × (a+b) / (c+d). The calculation for the expected values takes account of the size of the two corpora, so there is no need to normalise the figures before applying the formula. The log-likelihood value is then generated according to this formula[5]:

$$LL = 2\sum_i O_i \ln\left(\frac{O_i}{E_i}\right)$$

This equates to calculating LL as follows: LL = 2×((a×ln(a/E1)) + (b×ln(b/E2))). The word frequency list is then sorted by the resulting LL values. Sorting the list gives the effect of placing the largest LL value at the top of the list representing the word that has the most significant relative frequency difference between the two corpora. In this way, the words most indicative (or distinctive) of one corpus, as compared to the other corpus, occur at the top of the list. These are the same as key words as in Scott’s statistical sense.

The words that appear with roughly similar relative frequencies in the two corpora appear lower down the list.

The next stage in the proposed method is to carry out the same comparison at the POS and semantic level by deriving frequency lists of POS and semantic domain tags and applying the LL calculation to each entry in each of the two further lists. The resulting lists are again sorted on the LL value. Now, the POS tags and semantic domains that are most distinctive in one corpus as compared to the other appear at the top of the lists. By comparing the relative frequencies of the tags in each corpus, the overused and underused POS tags and semantic domains emerge in the same way that key words do in WordSmith.

It is at this point that the researcher must intervene and qualitatively examine concordance examples of the significant words, POS and semantic domains highlighted by this technique. I am not proposing a completely automated approach. Granger (1993) warns that we should not limit corpus investigation to what the computer can do for us automatically, and she quotes other authors who have come to the same conclusion. Woods et al (1986: 130) note that it is unsatisfactory simply to state that something was significant at the 1% level or was not significant at the 5% level, and Kretzschmar et al (1997) agree that the "analyst is always responsible for explanations" following a statistically significant result. The first aim of the POS and semantic comparisons is to reduce the number of key categories that the researcher should examine (addressing the criticism of the key words approach by Berber Sardinha, 1999). This qualitative examination is described by Leech and Fallon (1992) in more detail as the second stage

of their two stage process. The method described here could be substituted as an improvement on their stage one process. The inclusion of POS and semantic annotation also goes some way to addressing the question of sense distinctions described in Leech and Fallon (1992) and Baker (2004a). The sense distinctions marked by USAS are coarse-grained and may not match those required in specific studies, so care must be taken in interpreting the results. When carrying out the manual analysis of the results the researcher should also take account of possible tagging errors. Since the two taggers are automatic, there will be some mistakes in the process. Error rates quoted for the POS tagger are 96-97% (Leech and Smith 2000) and 91% for the semantic tagger (Rayson et al 2004b). Further discussion of this appears in Section 4.

As with applying the key words method, it is important to consider the issues relevant to comparison of corpora such as representativeness, homogeneity and comparability. For example, if we chose to compare a written corpus with a spoken corpus, it is very likely that lexical and grammatical differences between the spoken and written language will be exposed as well as differences in domain or content that we may wish to focus on.

To place the proposed method in the context of the Type III corpus linguistics methodology introduced in Section 1, it corresponds to Steps 3 and 4 (question and retrieve), assuming that Steps 1 and 2 (build and annotate) have already taken place. Step 5 (interpret) is perhaps the most important stage, and it corresponds to Leech and Fallon's stage two.

In order to provide software support for the method proposed here, I have implemented a web-based tool, Wmatrix.[6] Users of the tool can upload corpus texts via a web browser, have the texts automatically POS and semantically tagged using a *Tag Wizard*. Keyness comparisons at the word, POS and semantic domain level can be produced from the frequency profiles generated by the tool. Basic support is provided for concordancing and exporting of data. In work reported elsewhere in collaboration with colleagues, I have applied the keyness method in Wmatrix at the word level for social differentiation in the use of English vocabulary (Rayson et al 1997), at the POS level for profiling of learner English (Granger and Rayson 1998) and finally at the semantic domain level for analysis of technical documents from the software engineering domain (Sawyer et al 2005), the analysis of the concept of 'love' in Shakespeare's comedies and tragedies (Archer et al forthcoming) and analysis of interview transcripts for a study of knowledge transfer (Lockett 2006). The Wmatrix software is not the main focus of this paper but the tool has been used in the following case study and screenshots will be included to illustrate the output.

4. Case study

In order to demonstrate the differences in key items extracted at each of the three levels (word, part-of-speech and semantic domain), I have undertaken a comparison of the UK 2001 General Election manifestos of the Labour and Liberal Democratic (henceforth LibDem) parties. In a Type I or II study, I would decide before looking at the corpus data what phenomena I wish to investigate. I would then collect the necessary data and

examine the differences in the two manifestos to confirm or reject the hypothesis. In a Type III study as pursued here, I will examine the corpus data and let the analysis direct me to suggest further items to study. Though the documents are of relatively small size, this case study will still be able to show the relative merits of key words and key semantic domains. The Labour and LibDem 2001 General Election manifestos were downloaded from their respective websites.[7] The LibDem manifesto was available as a 248Kb Microsoft Word document (text only, no pictures) containing 57 pages and 20,344 words. This was converted to plain text (HTML format) using Microsoft Word. The Labour manifesto was available as four Adobe PDF files totalling 2.6Mb, the first and last of which represented the front and back covers of the document containing pictures and one short paragraph of text. The two remaining documents contained 44 pages of pictures and text. These were converted to plain text (HTML format) using Adobe Acrobat. The resulting conflated file contained 28,033 words.

4.1 Comparison at the word level

The analysis is begun by producing a word frequency list for the two corpora. The word frequency list for the Labour manifesto has over 4,200 entries, and the LibDem has over 3,600. In Table 2, I show the top 20 items in each list. The contents of the table illustrate four significant problems with using and comparing basic word frequency lists. These four problems highlight the need for the key words approach. Even though this technique is already well known, it is worth reiterating the issues here using the example data.

Table 2. 20 most frequent words in Labour and LibDem manifestos

	LibDem Manifesto		Labour Manifesto	
Rank	**Word**	**Frequency**	**Word**	**Frequency**
1	the	1174	the	1482
2	and	785	to	1112
3	to	736	and	1091
4	of	632	of	715
5	will	461	we	669
6	we	428	in	545
7	a	345	will	515
8	in	319	a	503
9	for	308	for	490
10	by	196	is	330
11	on	166	our	271
12	are	128	with	241
13	that	123	are	226
14	is	119	have	209
15	be	109	by	194
16	more	107	on	185
17	with	107	be	173
18	have	97	new	165
19	this	94	more	162
20	their	93	people	160

First, the frequencies cannot be compared directly unless they are normalised. The Labour manifesto contains nearly 8,000 more words than the LibDem one, so one would expect on average the frequencies to be higher for each word. Normalising the frequencies with respect to the corpus size means converting the frequency to a percentage value, or a value per thousand (or per million) words. Consider the word *will* in the table; it has a frequency of 515 in the Labour manifesto and 461 in the LibDem one, incorrectly suggesting higher usage by Labour. These observed figures should not be compared directly since the normalised values 1.84% for Labour and 2.27% for LibDem show that *will* occurs with greater relative frequency in the LibDem data. This difference is significant ($p < 0.005$ with LL value of 10.65 at 1 degree of freedom (d.f.)). It is worth repeating at this point that the LL calculation does include normalisation as part of the expected value formula.

Second, the high frequency words at the top of any word frequency list are generally of no further interest to anyone trying to differentiate the content of two corpora. The top twenty items usually consist of closed class words, such as articles (*the*), prepositions (*to*, *of*, *in*, *for* etc), conjunctions (*and*), and auxiliary verbs (*are*, *is*, *be*, *have*). At the bottom of the top twenty items in the Labour list, 'interesting' words from open classes appear which are worthy of further consideration such as the adjective *new*, the noun *people* and the adverb *more*. Despite this, high frequency words are of interest to some (Sinclair, 1999) and do lead to important findings (McEnery 2005: 170) so they should not be omitted automatically, for example by using stop-list filters.

Third, comparing the ranking of words is also misleading. The LibDem list places *more* three places higher up the list than its rank in the Labour list. Compare the relative frequencies of the word *more* in the two texts: Labour usage of 162 (0.58%) is higher than LibDem usage at 107 (0.53%). In fact, the difference is not significant (Log-likelihood value of 0.57 at 1d.f.), but one might be tempted to jump to the wrong conclusion given their relative positions in these lists.

Fourth, multiword expressions (Sag et al 2002) are not counted together. These have been referred to under various names, sometimes called lexical bundles (Biber et al 2004), fixed expressions and idioms (Moon 1998) and formulaic sequences (Schmitt 2004, Wray 2002). Depending on the purpose of the study, multiword expressions may be quite significant. For example, *to* in the LibDem data occurs 736 times. It can also occur in multiword prepositions, for example *subject to*, *according to* and *due to*.

Applying the keyness method at the word level, the relative use of words between Labour and LibDem manifestos can be compared. For 1 d.f., at 99% confidence (or $p < 0.01$), the cut-off of 6.63 would indicate that there are 283 words significantly overused or underused between the Labour and LibDem data. This reduces to 66 words significantly overused or underused at the 99.99% ($p < 0.0001$) level with the critical value 15.13, as recommended by an evaluation reported elsewhere (Rayson et al 2004a).

The Wmatrix tool illustrates the relative frequency differences using a *key word cloud*. This enables the user to visualise the key words in a similar manner to tag clouds employed in social networking web sites such as Flickr[8] and Delicious.[9] In those web

sites, an alphabetically sorted list of words (confusingly for this context called *tags*) are shown in a larger font if they are (manually) assigned more frequently to shared digital photographs (Flickr) or web site bookmarks (Delicious). However, in a key word cloud produced by Wmatrix the larger font indicates greater keyness. The top 100 key words produced by comparing the LibDem manifesto to the Labour text are shown in Figure 1.

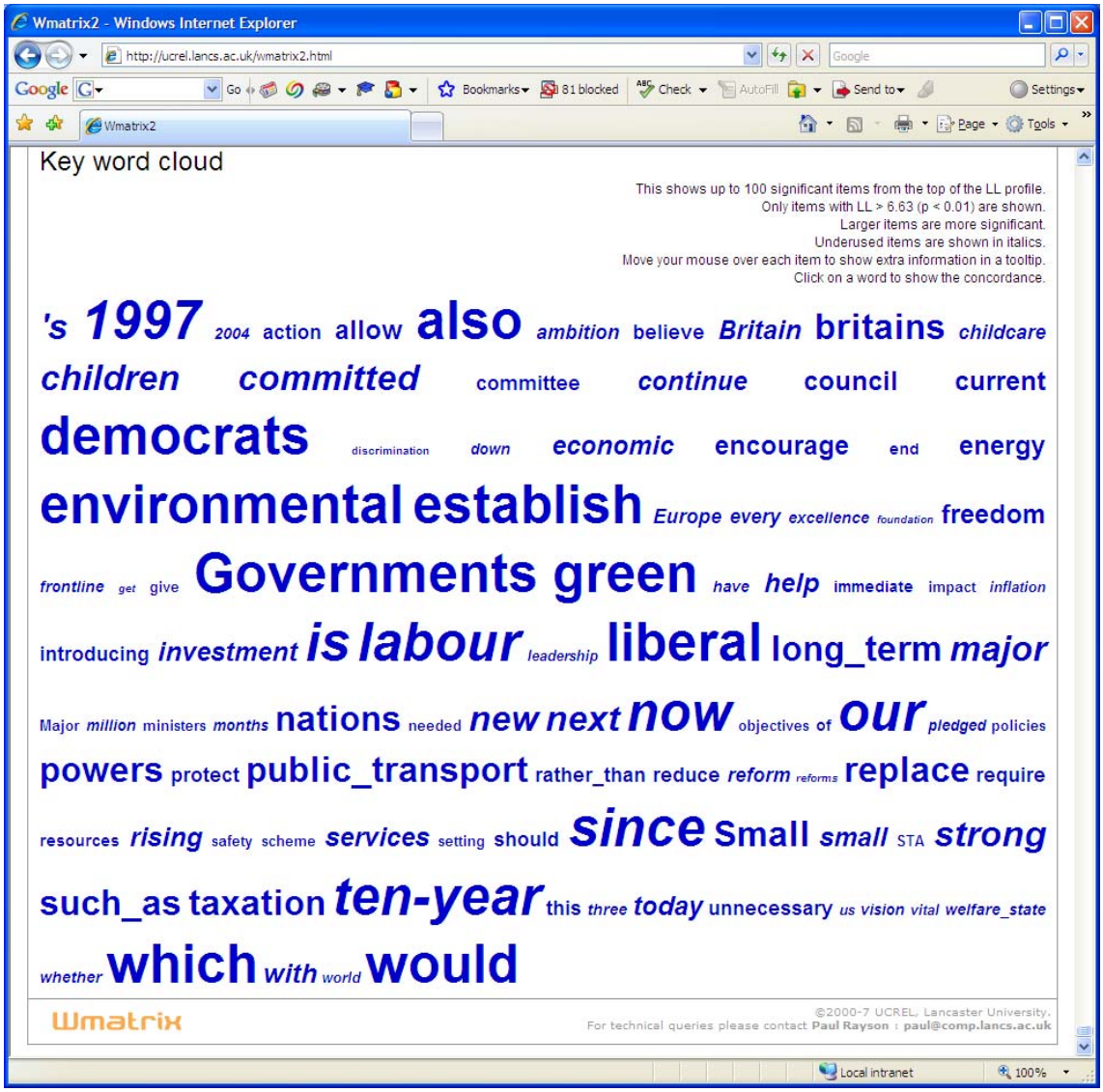

Figure 1. Wmatrix key word cloud of LibDem manifesto in comparison with Labour manifesto

The top twenty words (with the largest LL values) in this set are shown in Table 3. The table shows for each manifesto the frequency and relative frequency for each word in

the top twenty. The penultimate column indicates overuse (+) and underuse (-) of the word in the LibDem corpus with respect to the Labour corpus. In Figure 1, underused words are shown in italics.

Table 3. 20 most significant differences at word level between Labour and LibDem manifestos

		LibDem Manifesto		Labour Manifesto			
Rank	**Word**	**Freq.**	**Rel. Freq.**	**Freq.**	**Rel. Freq.**	**Overuse or underuse**	**LL**
1	liberal	47	0.23	0	0.00	+	81.43
2	would	70	0.34	10	0.04	+	71.90
3	democrats	40	0.20	0	0.00	+	69.30
4	our	76	0.37	271	0.97	-	62.59
5	labour	33	0.16	152	0.54	-	49.54
6	is	119	0.58	330	1.18	-	47.01
7	which	92	0.45	37	0.13	+	45.15
8	now	8	0.04	76	0.27	-	43.96
9	1997	4	0.02	54	0.19	-	36.75
10	green	26	0.13	2	0.01	+	32.82
11	environmental	47	0.23	14	0.05	+	30.99
12	establish	34	0.17	7	0.02	+	29.07
13	since	2	0.01	38	0.14	-	29.05
14	ten-year	0	0.00	25	0.09	-	27.28
15	also	88	0.43	50	0.18	+	26.31

16	Governments	15	0.07	0	0.00	+	25.99
17	britains	15	0.07	0	0.00	+	25.99
18	long_term	15	0.07	0	0.00	+	25.99
19	new	57	0.28	165	0.59	-	25.89
20	's	29	0.14	106	0.38	-	25.45

The first, third and fifth entries are unsurprising given that they show the names of the political parties. Looking at the concordance for liberal, there are 44 occurrences of Liberal Democrat(s) in the LibDem manifesto and none in the Labour one. There are some 33 references to the Labour party in the LibDem manifesto, although it has a lower frequency relative to the Labour document. It is therefore worth noting that the Labour manifesto chooses not to mention the Liberal Democrats at all.

The second most significant difference, with LL value of 71.90, alerts us to the fact that the word *would* is used almost 9 times relatively more frequently (0.04% compared to 0.34%) in the LibDem data. At this point I used the Wmatrix tool to look at a concordance of the key word *would* and this is shown in Figure 2.

Figure 2. Concordance of key word *would* from LibDem manifesto

Would, unlike *will*, denotes hypothetical or unreal events. Given the overuse of *would*, I might hypothesise that that the LibDem manifesto includes more mentions than the Labour one about possible future plans rather than definite plans, in other words, the LibDems, unlike Labour, did not expect to win the 2001 General Election. Moreover, looking lower down the word level comparison reminds us that *will* is also overused by the LibDems (2.27%) relative to Labour (1.84%) with a LL value of 10.65, which is

significant at 99% ($p < 0.01$). I will look at the relative use of modal verbs in the next section.

The fourth most significant difference is the word *our*, which is used significantly more in the Labour manifesto (0.97%) than in the LibDem statement (0.37%). In order to take this analysis further, the next step is to look at concordance lines for *our* in the two documents and initially classify the occurrences into those which refer to

- the British/English nation or people, e.g. "our children", "our sense of fair play"
- the Labour party/government, e.g. "our pledge not to extend VAT", "our reforms since 1997"
- ambiguous cases between the inclusive and exclusive classes, e.g. "incentives to meet our ambitions"

The relative use of these three categories might allow us to investigate whether Labour is intentionally using ambiguous language to make the reader feel that the party shares the same goals as they do. This mirrors the investigation of how collective identities are constructed through the use of inclusive and exclusive *we* in the language of New Labour, see Fairclough (2000: 35).

At the ninth position in the table is the number *1997* which is more frequent in the Labour manifesto (0.19% compared to 0.02%). This is unsurprising since 1997 was the year of the previous Labour victory in the General Election and the contexts for this key word show the manifesto detailing Labour's record in office since 1997. Labour's

achievements (since 1997) are also flagged by the key word *now*, which shows the eighth most significant difference, and is used over six times more frequently in the Labour text (0.27% compared to 0.04%). Figure 3 displays a section of the concordance lines for the key word *now* showing this trend.

```
NT color="#000000"> Europe . Britain now has the best combination </FONT> </Su
rld 's first University for Industry now offers over 400 skills courses . For
. Safer train protection systems are now being installed and will be extended
e been scrapped ; all new roads must now be strictly appraised for maximum ben
er ten years . &#163; 8.4 billion is now being invested in local authority sch
 are increasing , and over 100 towns now have bus services linked to train sta
ght historic wrongs . Every employee now has the right to four weeks ' paid ho
RDAs ) have been set up and why they now have extra money and new freedoms . <
00000"> to &#163; 1.7 billion a year now pledged to RDAs to </FONT> </P > <P
r cent of the national workforce are now employed in agriculture . But the ind
velopment priorities . CAP reform is now more possible ; Labour 's engagement
our platform &#8211; which is why we now have a unified grading scheme for hot
rt services ; and the Post Office is now obliged to prevent closure of rural p
s of coastal and inland flooding are now widely appreciated , and we are commi
ONT> </B> <B> <FONT color="#6C3C8A"> Now our ambition is for Britain to </FONT
c services are always second class . Now is the time to move forward . Economi
setting a clear national framework . Now we need to move on , empowering front
r refurbishment ; 20,000 schools are now connected to the internet ; there are
```

Figure 3. Concordance of key word *now* from Labour manifesto

At the nineteenth position in the table is the key word *new* which as one would expect is overused in the Labour manifesto. The slogan ‘New Labour, New Britain’ was first used at the 1994 Labour Party conference and Fairclough (2000: 18) discusses relevant themes such as renewal and modernisation.

4.2 Comparison at the POS level

In the previous section, I described four significant problems with using basic word frequency lists. As Barnbrook (1996: 53) writes, there are further limitations to the basic word frequency list related to the word forms as well as the frequencies. Inflected forms of words are not counted together, but word forms with two (or more) POS tags or meanings are counted together. This can be partially solved by annotating the text with POS tags and I used the CLAWS tagger (Garside and Smith, 1997) to assign word-class codes to the Labour and LibDem data.

Once the data has been tagged, we have access to what Francis and Kučera (1982) call 'grammatical words', i.e. words and their associated parts of speech. Examining the CLAWS tagged data, I found that the Labour and LibDem data contain no words that are ambiguous by POS. This means that each word in the data appears only within one part of speech, although in a much larger corpus (or corpus from another domain), you could find both noun and verb usage of the word *will* for example. I can compare the two files for their relative use of grammatical categories using the keyness method applied at the POS level. For $p < 0.01$, at 1 d.f. the cut-off of 6.63 would indicate that there are 35 POS tags significantly overused or underused between the Labour and LibDem data. At the 99.99% level ($p < 0.0001$), there are 17 significant POS tags. The top 20 tags (with the largest LL values) in this set are shown in Table 4.[10]

Table 4. 20 most significant differences at POS level between Labour and LibDem manifestos

		LibDem Manifesto		Labour Manifesto			
Rank	**POS**	**Freq.**	**Rel. Freq.**	**Freq.**	**Rel. Freq.**	**Overuse or underuse**	**LL**
1	MC	124	0.61	586	2.09	-	196.60
2	RT	13	0.06	105	0.37	-	55.25
3	VBZ	119	0.58	334	1.19	-	48.93
4	MD	22	0.11	122	0.44	-	48.13
5	NN2	1984	9.75	2246	8.01	+	40.47
6	DDQ	98	0.48	47	0.17	+	38.39
7	APPGE	199	0.98	438	1.56	-	31.58
8	VM	637	3.13	650	2.32	+	28.89
9	VV0	644	3.17	662	2.36	+	27.89
10	RR	379	1.86	369	1.32	+	22.48
11	GE	39	0.19	119	0.42	-	20.84
12	VH0	73	0.36	184	0.66	-	20.55
13	NNO	0	0.00	17	0.06	-	18.55
14	II21	68	0.33	41	0.15	+	18.20
15	IW	119	0.58	257	0.92	-	17.23
16	VVN	346	1.70	624	2.23	-	16.50
17	CSW	0	0.00	15	0.05	-	16.37
18	IO	633	3.11	714	2.55	+	13.37
19	NPM1	0	0.00	11	0.04	-	12.00

20	VVI	1043	5.13	1247	4.45	+	11.39

The most significant difference at the POS level is for the tag *MC* that marks cardinal numbers. The Labour manifesto includes more than three times as many cardinal numbers as in the LibDem one. This is largely the year 1997 as highlighted by the comparison at the word level. Also, *three*, *2004*, *2010* occur relatively frequently. Looking at concordances for these items, it can be observed that the Labour manifesto includes a large number of pledges for completion over the next three years, by 2004, or by 2010.

The second most significant difference at POS level is for the tag *RT* (time adverb) that includes occurrences of *now* and *today* more frequently in the Labour manifesto. I have already commented on the key word *now* above, but the key word *today* also seems to act as a marker for mentions of Labour's achievements since the previous election.

The key POS tag *APPGE* (pre-nominal possessive pronoun) is overused in the Labour text (LL value of 31.58) and this is mostly due to the preference in the Labour manifesto for the key word *our* as described above.

With a LL value of 28.89, the LibDem manifesto overuses modal verbs (VM). This word class includes *will* and *would* and I have already discussed these key words from the word level comparison. However, I can now examine the comparative frequency of use of the other modal verbs. This is illustrated by Figure 4 which shows the LibDem's

preference for *would*, *could* and *may*, and conversely Labour's preference for *can*, *must*, *might* and *shall*.

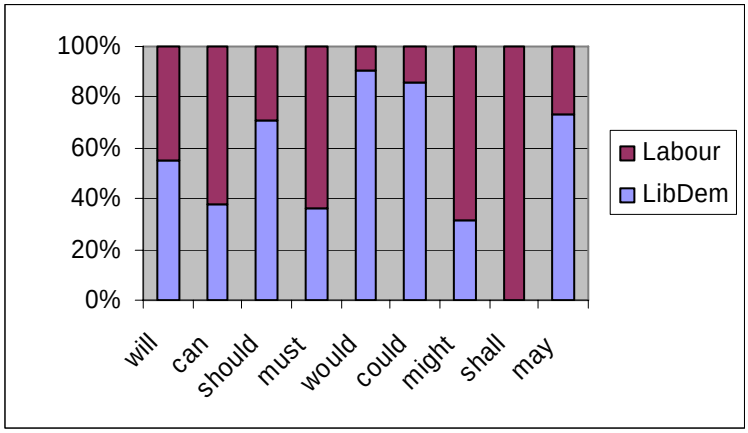

Figure 4. Relative use of modal verbs in LibDem and Labour manifestos

4.3 Comparison at the semantic level

I used the USAS tagger described in Rayson et al (2004b) to assign semantic field (domain) tags to the Labour and LibDem data. I can then compare the two resulting files for their relative use of USAS categories using the keyness method applied at the semantic level. For $p < 0.01$ with 1 d.f., the cut-off of 6.63 would indicate that there are 65 USAS tags significantly overused or underused between the Labour and LibDem data. At the $p < 0.0001$ level, the critical value is 15.13, giving 23 significant USAS tags. The top 20 tags (with the largest LL values) in this set are shown in Table 5[11] and the Wmatrix visualisation of the resulting *key domain cloud* is shown as Figure 5.

Table 5. 20 most significant differences at semantic level between Labour and LibDem manifestos

		LibDem Manifesto		Labour Manifesto				
Rank	**USAS tag**	**Freq.**	**Rel. Freq.**	**Freq.**	**Rel. Freq.**	**Overuse or underuse**	**LL**	**Semantic domain**
1	N1	141	0.69	554	1.98	-	147.78	Numbers
2	A1.7-	116	0.57	34	0.12	+	77.51	No constraint
3	G1.1	377	1.85	297	1.06	+	52.41	Government
4	W5	107	0.53	45	0.16	+	49.81	Green issues
5	Z3	208	1.02	146	0.52	+	39.85	Other proper names
6	M3	148	0.73	96	0.34	+	34.08	Vehicles and transport on land
7	A3+	253	1.24	521	1.86	-	28.63	Existing
8	S6-	48	0.24	17	0.06	+	27.00	No obligation or necessity
9	S7.4+	80	0.39	43	0.15	+	26.31	Allowed
10	T2-	74	0.36	39	0.14	+	25.13	Time: Ending

11	N5	73	0.36	193	0.69	-	24.47	Quantities
12	O4.3	32	0.16	8	0.03	+	24.14	Colour and colour patterns
13	I3.1	155	0.76	332	1.18	-	21.54	Work and employment
14	T3-	106	0.52	246	0.88	-	21.37	Time: New and young
15	X2.4	88	0.43	57	0.20	+	20.33	Investigate, examine, test, search
16	A2.1+	147	0.72	310	1.11	-	18.87	Change
17	N4	45	0.22	124	0.44	-	17.40	Linear order
18	A6.1-	85	0.42	59	0.21	+	16.74	Comparing: Different
19	N5-	97	0.48	71	0.25	+	16.67	Quantities: little
20	A5.1+++	20	0.10	70	0.25	-	15.69	Evaluation: Good

The most significant difference (LL value 147.78) in the semantic comparison is for the tag *N1* representing the semantic field *numbers*. This is largely due to words with POS tag MC (as highlighted by the POS level comparison) being overused in the Labour manifesto.

The ninth most significant difference (LL value 77.51) indicates the overuse of the semantic domain of *Allowed* (S7.4+) in the LibDem manifesto. Upon examining the concordance for this tag (part of which is shown in Figure 6), it can be seen that 47 of the entries are the word *liberal*, and 44 of these refer to the Liberal Democrat(s). In fact, these items are mistagged by the automatic semantic tagger and should obtain the G1.2 tag indicating the *political* semantic field. When I recalculate by omitting the 44 mistakes, the relative frequencies are 0.43% in the LibDem document compared to 0.17% in the Labour one, and this still results in a significant LL value of 28.36. Looking at the other terms in this field such as *allow*, *right*, and *entitled*, I might form the hypothesis that the LibDem manifesto focuses more on personal freedoms than the Labour text, and study this aspect in more detail. This hypothesis is corroborated by evidence from the second most significant difference, which is the domain of *No constraint* (A1.7-) overused in the LibDem manifesto (0.57% compared to 0.12%). The minus sign at the end of the tag indicates the negative end of the *Constraint* (A1.7) domain and the words I find within this category are *freedom(s)* and *liberties*.

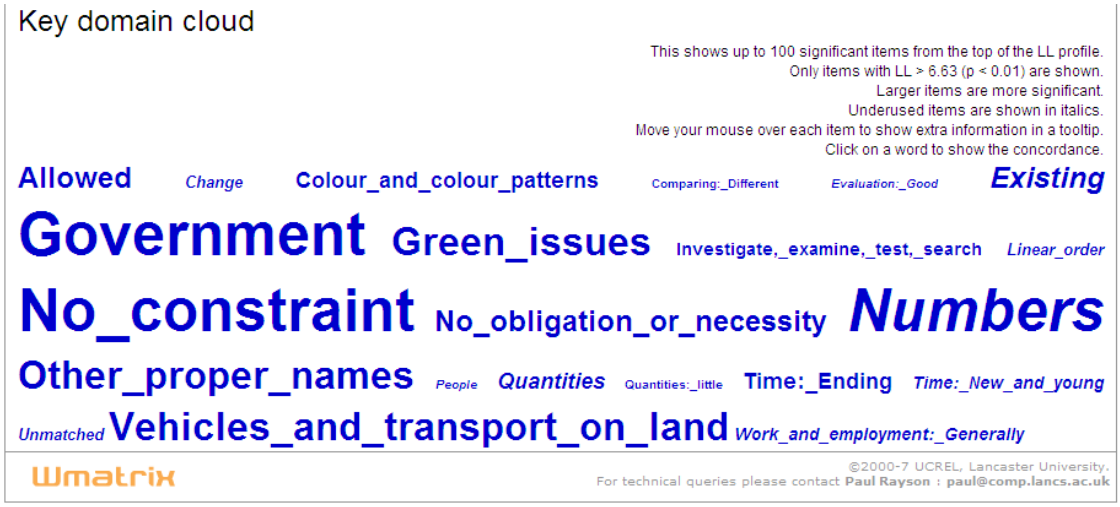

Figure 5. Key domain cloud for the Labour and LibDem manifesto comparison

The fourteenth most significant category is *Time: new and young* (T3-) which is overused in the Labour manifesto (0.88%) relative to LibDem (0.52%). This category marks the words *new*, *child(ren)*, *young*, and *modern* amongst others. The key word *new* has already been identified by the word level comparison. The young/family terms relate to the family policy area mentioned below. A related category at position sixteen is that of *Change* (A2.1+), which is overused in the Labour document (1.11%) compared to the LibDem text (0.72%). This category contains words such as *reform(s)*, *develop(ment)* and *change*. Fairclough (2000: 18) links reform to the sense of political renewal conveyed by Labour indicated by key words such as *new*.

```
n: yes"> </span> We will also allow     people to stand for elected of
"> </span> We will extend the right     to vote by post and investigat
wers of Select Committees and allow     more pre-legislative scrutiny
s more say over the budget by allowing  them to propose spending amend
te from the Finance Bill , to allow     for greater consultation on ta
acerun: yes"> </span> We will allow     the Welsh Assembly the right t
 allow the Welsh Assembly the right     to pass primary legislation an
cerun: yes"> </span> We would allow     further devolution of powers a
span> They are essential to a liberal   society in which people are en
black;layout-grid-mode:line'> Liberal   Democrats will : <o:p> </o:p>
trong framework of individual rights    , extending the protection alr
by European law , so that the rights    of the individual outweigh the
d personal relationship legal rights    , such as next-of-kin arrangem
span style='color:black'> The Right     to Know and the Right to Priva
k'> The Right to Know and the Right     to Privacy <o:p> </o:p> </span
e individuals should have the right     to know as much as possible ab
eplace the system of warrants approved  by Ministers with a system of
```

```
by Ministers with a system of approval  by judges to remove any confli
r:black;font-style:normal'> A Right     to Environmental Information ,
:normal'> We will protect the right     to legal and peaceful protest
e that farm animals should be entitled  to high welfare standards . <s
```

Figure 6. Concordance of key domain *allowed* from LibDem manifesto

At eighteenth position is the domain *comparing: different* (A6.1-) which is used to a greater extent in the LibDem manifesto (0.42% compared to 0.21%). This includes words such as *other(s)*, *discrimination*, *different*, *separate*, and *conflict*. The reasons behind this difference are not clear and require further investigation. However, I might hypothesise that the lower count in the Labour text stems from the 'one-nation politics' of Labour whose discourse is inclusive and consensual, and would de-emphasize these words which have negative connotations.

The tenth entry for *time: ending* (T2-) can be examined together with *time: beginning* (T2++) which occurs just off the bottom of the table at position twenty-three. Continuity domains occur more frequently in the Labour document and the reverse is true for the domain of ending/stopping. From the concordances of these domains they seem to mark government policies that Labour would continue pursuing if they were to stay in power and that the Liberal Democrats would end if they were elected.

Emerging from the comparison at the semantic or domain level are relative differences in the prominence of party policy areas. Labour's document focuses more on *work and employment* (USAS tag I3.1), and *kin* (S4) representing family issues. The LibDem manifesto devotes more of its content to *vehicles and transport* (M3) reflecting transport policy, and to *green issues* (W5) and *colour* (O4.3) indicating

green/environmental policy. This is also shown at position 25 just outside the table where *substances and materials* (O1) is used to a greater extent in the LibDem manifesto (0.14% relative to 0.04%). This category includes words such as *fuel(s)*, *air*, *water*, *gas*, and *petrol*. The use of these words seems to be partly related to the LibDem textual focus on environmental issues including mentions of fuel taxation policy and conservation of resources. In addition, at the word level in Table 3 with the key words *green* and *environmental* showing increased use in the LibDem document, but the comparison at semantic level provides more reliable evidence of the observed focus since several key words and phrases, e.g. *pollution* and *environmentally friendly*, contribute and confirm it.

4.4 Summary of the worked example

The ability to extract key semantic domains and create or suggest hypotheses about major trends from the two documents demonstrates clearly the advantages of the comparison at the semantic level in addition to the (stylistic comparison) at the word and POS levels. I had to examine a much smaller number of key domains in the semantic comparison than the number of key words.[12] Therefore overall trends are easier to identify. Furthermore, it is not possible to identify some of the significant semantic domains at the word or POS level. Consider, *work and employment* (I3.1) mentioned above. Of the words in this category such as *work(ing)*, *staff*, *(un)employment*, *job(s)*, and *employees*, only the word *work* (LL 10.60) is significant in the word level comparison. Collecting together words into their semantic fields allows us to see trends that are invisible at the word level. Henry and Roseberry (2001: 101)

also report a similar finding where an important semantic class groups together low frequency words that would otherwise have been missed.

Two further advantages of the comparison at the semantic level are that multiword expressions are counted together and variants within a lemma are usually grouped together. Multiword expressions are identified by the list of flexible templates associated with the USAS system as described in Piao et al (2005). In the LibDem data the following terms are identified amongst others: *local authorities*, *public transport*, *human rights*, *United Kingdom*, *league tables*. In the absence of information to the contrary, the USAS system groups variants within a lemma by using stemming and lemmatisation rules for dictionary look-up.

In this section, I have looked at the language used in the United Kingdom General Election manifestos of the Labour and Liberal Democratic (LibDem) parties from the June 7th 2001 election. The initial results have suggested numerous avenues for further investigation to pursue a Type III (data-driven) study, ranging from lexical studies, through grammatical variation to analyses of party political differences (political linguistics). Some of these avenues for investigation are summarised here:

1. The inclusive language of Labour is indicated by the greater use of the word *our* in their manifesto.
2. The differing use of modal verbs is found between the LibDem and Labour manifestos, signposted by the overuse of *would* in the LibDem manifesto.

3. The differing use of permission and freedom domains is also found, highlighted by significantly greater use of these domains in the LibDem manifesto.
4. The political renewal senses are conveyed by overuse of terms such as *new*, *modern*, *reform*, and *change* in the Labour manifesto.
5. Party policy differences between LibDem and Labour are indicated by significant differences in the relative use of domains related to environmental issues, family issues, work and employment, and transport.

5. Summary and conclusions

In the introduction, I described the typical processes involved in corpus linguistics methodology. I drew a distinction between Type I (microscopic) studies, where research questions focus on specific linguistic features, and Type II (macroscopic) studies on characteristics of whole texts or varieties of language. I also sketched out a data-driven approach (classified as Type III) where research questions emerge from iterative analyses of corpus data. Elements of both the corpus-based and corpus-driven paradigms can be combined using this data-driven approach. Systematic approaches to the comparison of corpora were shown to support this data-driven method and, in particular, I focussed on the well-known key words method implemented in software such as WordSmith tools.

In order to address criticisms of the key words method by Berber Sardinha (1999), Gries (2006) and Baker (2004a), I proposed an extension of the key words method to key

parts-of-speech and key semantic domains as described in detail in this paper. I have utilised the Wmatrix software tool that has been implemented to support the proposed methodology in order to carry out frequency profiling of corpora and comparison of those profiles. In order to suggest possible research questions for further investigation, the proposed method uses the log-likelihood ratio statistic to compare frequencies and then rank them in terms of significance of differences.

A worked example illustrated the data-driven method with two corpora consisting of UK 2001 General Election manifestos. Key grammatical categories and key semantic domains are used to group together lower frequency words and multiword expressions which would, by themselves, not be identified as key, and would otherwise be overlooked. Comparison at the POS and semantic levels reduces the number of key items that the researcher should examine, thus addressing the practical problem with key words identified by Berber Sardinha (1999). The proposed method can replace stage one of Leech and Fallon's (1992) process (quantitative extraction), and it assists in their stage two (qualitative examination). The use of POS and semantic analysis addresses to some extent the limitations of the key words approach described by Gries (2006) and the need for sense distinctions identified by Baker (2004a).

Currently, the method described here has been applied only to English language corpora although additional semantic taggers have been developed for Finnish (Löfberg et al 2005) and Russian (Mudraya et al 2006) which may permit its application to those languages. The method itself is automatic, but I am not proposing a completely automatic procedure. Some thought is required in choosing an appropriate reference

corpus when using the method to compare one corpus against a much larger representative corpus. Careful manual analysis of concordances of key words and key domains is obviously required to check for mistagging and poor dispersion of high frequency items. In future work, we may need to take account of effect sizes, an issue highlighted by Gries (2006).

There are two main contributions contained in this paper. First is the presentation of the complete method for extending the keyness technique from key words to key semantic domains. This has been shown as a way to combine elements from both the corpus-based and corpus-driven paradigms within corpus linguistics. Secondly, this paper brings together the three levels of analysis in one comparative case study showing the differences observed at each of the three levels. The method described in this paper and the supporting Wmatrix software have been applied in a growing number of studies and I hope that further research will show that they are applicable across as wide a spectrum as the key words method on which they are based.

Acknowledgements

I wish to thank Dawn Archer, Geoffrey Leech and Roger Garside for helpful discussions during the development of this work. I would also like to thank the editor of this journal, Michaela Mahlberg, and two anonymous reviewers for their valuable comments and suggestions.

Notes

[1] Also see the papers listed on the WordSmith website: http://www.lexically.net/wordsmith/corpus_linguistics_links/papers_using_wordsmith.htm
[2] http://www.methodsnetwork.ac.uk/activities/es01mainpage.html
[3] http://www.disas.unisi.it/keyness/index.php
[4] Further details of the software and tagsets employed can be found at http://ucrel.lancs.ac.uk/claws/ and http://ucrel.lancs.ac.uk/usas/
[5] This simpler version of the LL formula comes from Read and Cressie (1988: 3) rather than Dunning (1993) for example
[6] For more details including a tutorial, see http://ucrel.lancs.ac.uk/wmatrix/
[7] The Labour Party website is http://www.labour.org.uk/ and the Liberal Democrat Party website is http://www.libdems.org.uk/ - it should be noted that the 2001 manifestos are no longer available from these websites. The manifestos will be made available at the author's website prior to publication
[8] http://www.flickr.com/photos/tags/
[9] http://del.icio.us/tag/
[10] A full list of CLAWS C7 tagset can be found at http://ucrel.lancs.ac.uk/claws7tags.html
[11] A full list of USAS tags can be found at http://ucrel.lancs.ac.uk/usas/USASSemanticTagset.pdf
[12] Nevill-Manning et al (1999) similarly report the speed improvements for finding useful information in large collections (digital libraries) using a hierarchical structure of phrases.

References

Archer, D. (ed.) (forthcoming) *What's in a word-list? Investigating word frequency and keyword extraction*. Aldershot: Ashgate.

Archer, D., Culpeper, J. and Rayson, P. (forthcoming) Love - a familiar or a devil? An exploration of key domains in Shakespeare' Comedies and Tragedies. In Archer, D. (ed.) *What's in a word-list? Investigating word frequency and keyword extraction*. Aldershot: Ashgate.

Baker, P. (2004a) Querying keywords: questions of difference, frequency and sense in keywords analysis. *Journal of English Linguistics*. 32: 4, pp. 346-359.

Baker, P. (2004b) 'Unnatural Acts': Discourses of homosexuality within the House of Lords debates on gay male law reform. *Journal of Sociolinguistics* 8 (1), 88–106.

Barnbrook, G. (1996). *Language and Computers: a practical introduction to the computer analysis of language*. Edinburgh: Edinburgh University Press.

Berber Sardinha, T. (1999). Using KeyWords in text analysis: Practical aspects. *DIRECT Working Papers* 42, São Paulo and Liverpool. Available online http://www2.lael.pucsp.br/direct/DirectPapers42.pdf

Biber, D. (1988). *Variation across speech and writing.* Cambridge: Cambridge University Press.

Biber, D. (1995). *Dimensions of register variation : a cross-linguistic comparison*. Cambridge: Cambridge University Press.

Biber, D. and Finegan, E. (1989). Drift and the evolution of English style: a history of three genres. *Language* 65, pp. 487 – 517.

Biber, D., Conrad, S. and Cortes, V. (2004) If you look at ...: Lexical Bundles in University Teaching and Textbooks. *Applied Linguistics* 25(3):371-405.

Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19 (1), pp. 61-74.

Fairclough, N. (2000). *New Labour, New Language?* London: Routledge.

Firth, J. R. (1935). The technique of semantics. *Transactions of the Philological Society*, London: David Nutt, pp. 36 – 72.

Francis, G. (1993). A corpus-driven approach to grammar: principles, methods and examples. In Baker, M., Francis, G., and Tognini-Bonelli, E. (eds.) *Text and technology: in honour of John Sinclair.* Amsterdam: John Benjamins, pp. 137 – 156.

Francis, W. N. and Kučera, H. (1982). *Frequency Analysis of English Usage: Lexicon and Grammar*. Boston: Houghton Mifflin.

Garside, R. and Smith, N. (1997). A Hybrid Grammatical Tagger: CLAWS4. In Garside, R., Leech, G., and McEnery, A. (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*, London: Longman, pp. 102 – 121.

Granger, S. (1993). International Corpus of Learner English. In Aarts, J., de Haan, P., and Oostdijk, N. (eds.) *English language corpora: Design, analysis and exploitation*. Amsterdam: Rodopi, pp. 57 – 71.

Granger, S. (1999). Use of tenses by advanced EFL learners: evidence from an error-tagged computer corpus. In Hasselgård, H. and Oksefjell, S. (eds.) *Out of corpora: studies in honour of Stig Johansson*. Amsterdam, Rodopi. pp. 191 – 202.

Granger, S. and Rayson, P. (1998). Automatic profiling of learner texts. In S. Granger (ed.) *Learner English on Computer*. London: Longman, pp. 119 – 131.

Gries, S. T. (2006). Exploring variability within and between corpora: some methodological considerations. *Corpora* 1(2), pp. 109-151.

Henry, A. and Roseberry, R. L. (2001). Using a small corpus to obtain data for teaching a genre. In Ghadessy, M., Henry, A. and Roseberry, R. L. (eds.) *Small corpus studies and ELT: theory and practice*, Amsterdam: John Benjamins, pp. 93 – 133.

Hoffmann, S. and Lehmann, H. M. (2000). Collocational evidence from the British National Corpus. In Kirk, J. M. (ed.) *Corpora galore: analyses and techniques in describing English*. Amsterdam: Rodopi, pp. 17 – 32.

Hofland, K. and Johansson, S. (1982). *Word frequencies in British and American English*. Bergen: The Norwegian Computing Centre for the Humanities.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Kilgarriff, A. and Tugwell, D. (2002). Sketching words. In Corréard, M-H. (ed.) *Lexicography and natural language processing: a festschrift in honour of B. T. S. Atkins*, Euralex, pp. 125 – 137.

Kretzschmar, W. A., Meyer, C. F., and Ingegneri, D. (1997). Uses of inferential statistics in corpus studies. In Ljung, M. (ed.) *Corpus-based studies in English: papers from the seventeenth International Conference on English language research on computerized corpora (ICAME 17)*, Stockholm, May 15-19, 1996, Amsterdam: Rodopi, pp. 167 – 177.

Lee, D. Y. W. (2000). *Modelling variation in spoken and written language: the multi-dimensional approach revisited*. PhD thesis, Linguistics Department, Lancaster University.

Leech, G. (1992). Corpus linguistics and theories of linguistic performance. In Svartvik, J. (ed.) *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82*, Stockholm, 4 – 8 August 1991. Berlin: Mouton de Gruyter, pp. 105 – 122.

Leech, G. and Fallon, R. (1992). Computer corpora – what do they tell us about culture? *ICAME Journal*, 16, pp. 29 – 50.

Leech, G. and Smith, N. (2000). *Manual to accompany The British National Corpus (Version 2) with Improved Word-class Tagging*. (Accessed 20th December 2007) http://ucrel.lancs.ac.uk/bnc2/bnc2postag_manual.htm

Lockett, N. (2006). *InfoLab21 Knowledge Transfer Study*. (Accessed 14th December 2007). http://www.communitiesofinnovation.org/docs/InfoLab21KTStudyFinal.pdf

Löfberg, L., Piao, S., Rayson, P., Juntunen, J-P., Nykänen, A., and Varantola, K. (2005) A semantic tagger for the Finnish language. In *proceedings of the Corpus Linguistics 2005 conference*, July 14-17, Birmingham, UK. http://www.corpus.bham.ac.uk/PCLC/

Lyne, A. A. (1985). *The vocabulary of French business correspondence*. Geneva: Slatkine.

McEnery, T. (2005). *Swearing in English: Bad Language, Purity and Power from 1586 to the Present.* London: Routledge.

Mahlberg, M. (2007). Clusters, key clusters and local textual functions in Dickens. *Corpora* 2 (1), pp. 1-31.

Meyer, C. F. (1991). A corpus-based study of apposition in English. In Aijmer, K. and Altenberg, B. (eds.), *English Corpus Linguistics: Studies in honour of Jan Svartvik*. London: Longman, pp. 166 – 181.

Moon, R. (1998). *Fixed Expressions and Idioms in English: A Corpus-Based Approach.* Oxford: Oxford University Press.

Mudraya, O., Babych, B., Piao, S., Rayson, P., Wilson, A. (2006). Developing a Russian semantic tagger for automatic semantic annotation. In *proceedings of Corpus*

Linguistics 2006, St. Petersburg, Russia, 10-14 October 2006, pp. 290-297. http://ucrel.lancs.ac.uk/publications/MudrayaEtAlCL2006English.pdf

Nevill-Manning, C. G., Witten, I. H., and Paynter, G. W. (1999). Lexically-Generated Subject Hierarchies for Browsing Large Collections. *International Journal of Digital Libraries*, 2 (2/3), pp. 111 – 123.

Oakes, M.P. and Farrow, M. (2007) Use of the Chi-Squared Test to Examine Vocabulary Differences in English Language Corpora Representing Seven Different Countries, *Literary and Linguistic Computing* 22(1): 85-99.

Oakes, M. P. (forthcoming). Measures from information retrieval to find the words which are characteristic of a corpus. In *proceedings of Practical Applications in Language and Computers (PALC 2007) conference*. 19-22 April 2007, Lodz University, Poland, Frankfurt: Peter Lang.

Ooi, V. B. Y. (2000). Asian or Western realities? Collocations in Singaporean-Malaysian English. In Kirk, J. M. (ed.) *Corpora galore: analyses and techniques in describing English*. Amsterdam: Rodopi, pp. 73 – 89.

Piao, S., Rayson, P., Archer, D., McEnery, T. (2005) Comparing and combining a semantic tagger and a statistical tool for MWE extraction. *Computer Speech and Language*, 19 (4), pp. 378 - 397

Rayson, P. (2003). *Matrix: A statistical method and software tool for linguistic analysis through corpus comparison*. Ph.D. thesis, Lancaster University.

Rayson, P., Leech, G., and Hodges, M. (1997). Social differentiation in the use of English vocabulary: some analyses of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics*. 2 (1), pp 133 – 152.

Rayson P., Berridge D. and Francis B. (2004a). Extending the Cochran rule for the comparison of word frequencies between corpora. In Volume II of Purnelle G., Fairon C., Dister A. (eds.) Le poids des mots: Proceedings of the 7th International Conference on Statistical analysis of textual data (JADT 2004), Louvain-la-Neuve, Belgium, March 10-12, 2004. Louvain-la-Neuve: Presses universitaires de Louvain, pp. 926 - 936.

Rayson, P., Archer, D., Piao, S. L., McEnery, T. (2004b). The UCREL semantic analysis system. In proceedings of the workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks in association with 4th International Conference on Language Resources and Evaluation (LREC 2004), 25th May 2004, Lisbon, Portugal. Paris: European Language Resources Association, pp. 7-12.

Read, T. and Cressie, N. (1988). *Goodness of fit Statistics for Discrete Multivariate Data*. New York: Springer-Verlag.

Ringbom, H. (1998). High-frequency verbs in the ICLE corpus. In Renouf, A. (ed.) *Explorations in corpus linguistics*. Amsterdam: Rodopi, pp. 191 – 200.

Sag, I., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword expressions: a pain in the neck for NLP. In *proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics* (CICLing-2002), Mexico City, Mexico, pp. 1-15.

Sawyer, P., Rayson, P. and Cosh, K. (2005) Shallow Knowledge as an Aid to Deep Understanding in Early Phase Requirements Engineering. *IEEE Transactions on Software Engineering*. 31 (11), November, 2005, pp. 969 - 981.

Schmitt, N. (2004). *Formulaic Sequences: Acquisition, Processing, and Use*. Amsterdam: John Benjamins.

Scott, M. (1997). PC analysis of key words – and key key words. *System* 25 (2), pp. 233 – 245.

Scott, M. (2000a). Reverberations of an Echo. In Lewandowska-Tomaszczyk, B. and Melia, P. J. (eds.) *PALC'99: Practical applications in language corpora: papers from the International conference at the University of Łódź*, 15 – 18 April 1999. Frankfurt: Peter Lang, pp. 49 – 65.

Scott, M. (2000b). Focusing on the text and its key words. In Burnard, L. and McEnery, T. (eds.) *Rethinking language pedagogy from a corpus perspective: papers from the*

third international conference on teaching and language corpora. Frankfurt: Peter Lang, pp. 104 – 121.

Scott, M. (2001a). Comparing corpora and identifying key words, collocations, and frequency distributions through the WordSmith Tools suite of computer programs, in Ghadessy, M., Henry, A. and Roseberry, R. L. (eds.) *Small corpus studies and ELT: theory and practice*. Amsterdam: John Benjamins, pp. 47 – 67.

Scott, M. (2001b). Mapping key words to problem and solution. In Scott, M. and Thompson, G. (eds.) *Patterns of Text: in honour of Michael Hoey*, Amsterdam: John Benjamins, pp. 109 – 127.

Scott, M. and Tribble, C. (2006) *Textual Patterns: Key words and corpus analysis in language education*. Amsterdam: John Benjamins.

Sinclair, J. (1999). A way with common words. In Hasselgård, H. and Oksefjell, S. (eds.) *Out of corpora: studies in honour of Stig Johansson*. Amsterdam: Rodopi. pp. 157 – 179.

Stubbs, M. (1993). British traditions in text analysis: from Firth to Sinclair. In Baker, M., Francis, G., and Tognini-Bonelli, E. (eds.) *Text and technology: in honour of John Sinclair.* Amsterdam: John Benjamins, pp. 1 – 33.

Stubbs, M. (1996). *Text and corpus analysis: computer-assisted studies of language and culture*. Oxford: Blackwell.

Teubert, W. (2001). A province of a federal superstate, ruled by an unelected bureaucracy – keywords of the Euro-sceptic discourse in Britain. In Musolff, A., Good, C., Points, P., Wittlinger, R. (eds.) *Attitudes towards Europe: language in the unification process*. Aldershot: Ashgate, pp. 45 – 86.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam: John Benjamins.

Tribble, C. (2000). Genres, keywords, teaching: towards a pedagogic account of the language of project proposals. In Burnard, L. and McEnery, T. (eds.) *Rethinking language pedagogy from a corpus perspective: papers from the third international conference on teaching and language corpora*. Frankfurt: Peter Lang, pp. 75 – 90.

Tribble, C. (2001). Small corpora and teaching writing. In Ghadessy, M., Henry, A. and Roseberry, R. L. (eds.) *Small corpus studies and ELT: theory and practice*, John Amsterdam: John Benjamins, pp. 381 – 408.

Woods, A., Fletcher, P., and Hughes, A. (1986). *Statistics in language studies*. Cambridge: Cambridge University Press.

Wierzbicka, A. (1997). *Understanding cultures through their key words*. Oxford: Oxford University Press.

Williams, R. (1983). *Keywords: a vocabulary of culture and society*, 2nd edition, London: Fontana Press.

Wray, A. (2002). *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

Xiao, Z. and McEnery, T. (2005). Two Approaches to Genre Analysis: Three Genres in Modern American English. *Journal of English Linguistics*, Vol. 33, No. 1, 62-82.