Pergamon

System, Vol. 25, No. 2, pp. 233–245, 1997

PII: S0346-251X(97)00011-0
0346-251X/97 $17.00+0.00

PC ANALYSIS OF KEY WORDS - AND KEY KEY WORDS

MIKE SCOTT

University of Liverpool, Applied English Language Studies Unit, Department of English Language and Literature, Modern Languages Building, Liverpool L69 3BX

Some words are lonelier than others. The fewer derivatives and analogues a word has, the weaker the background of bodily habit, the lonelier it is. Anything may happen to it. (Firth, 1930 [in 1966: p. 183])

Taking as its starting point the category text (as opposed to clause, collocation span, sentence etc.) this paper proposes and illustrates a method of identifying key words in text, and leads from this to the notion of key key words (words that are key in many texts). A key key word is shown to have associates: words that are key in the same texts as a given key key word. Finally, these associates can be grouped together in clumps, which themselves are revealing about text schemata and stereotype. The number of clumps in which any key key word participates relates not to ambiguity but to a diversity of perceived roles; the members of the clump describe and illuminate these stereotypical roles. © 1997 Elsevier Science Ltd

INTRODUCTION

The term PC applies in two senses, for the present paper[1] concerns itself with an initial attempt to pin down certain key words in late 20th Century culture, a time when key cultural terms are being much questioned. The paper points to a procedure for clustering key words in culturally significant ways; in other words, clusters of associated key words can provide a representation of socially important concepts. The purpose is akin to that of Raymond Williams (1976)[2], though the procedure is quite different from his, being based on PC processing of large numbers of texts. As we shall see, there is stereotyping in the representation of socially important concepts. Indeed, without stereotyping the procedure to be described below would not work, for it is because key words are not equally associated with each other that it has proved possible to cluster them together in clumps.

The paper, then, is concerned with stereotype. This is really a wider and more emotionally-charged term for what Bartlett (1932), Rumelhart (1975) and others have termed schemata: the socially determined networks of links between ideas.[3] The classic example is the restaurant schema, which associates the ideas |menu|, |cook|, |waiter|, |payment|, etc.

STARTING POINTS

The view of Corpus Linguistics which underlies my current work is that the text is the central category, in preference to others:

Table 1. Potential Categories. √ = important to non-linguists; + = easy to determine in a corpus because there's an unambiguous marker belonging to a closed set

letter	√	+
morpheme		+
word	√	+
cluster of words		+
phrase (as in VP)		
collocation span		+
clause		
heading	√	+
sentence	√	+
paragraph	√	+
section	√	?
chapter	√	+
text	√	+
text-type	?	?

Although the list of categories in Table 1 is neither complete nor homogeneous, it serves to illustrate two ideas: that not all the items marked √ in the middle column are, or have been defined and used as, categories by the professional linguist, even though they are of some importance to the non-linguist;[4] second, that Corpus Linguistics has a number of alternative fairly easily machine-identified starting-points, which do not include all the main categories of General Linguistics, where some theories have had the sentence as their major category while others have chosen to erect clause as the major category.

A great deal of Corpus Linguistics work has started from the word and moved to the collocation span, in order to view the word in context. Note that the expression key word in context (KWIC) implies that context is a span of perhaps 80 characters or so. This ignores other senses of context which include at the very least a sense of "the story so far" or global context (Scott, 1991: p. 97 discusses a hierarchy of six levels of co-text).

Much of the work at Birmingham has been based on the word in context, by which is typically meant the word within a collocation span extending about four words on each side of a concordance node. This has led to a vast production, such as the Cobuild dictionaries and their latest publication on verb patterns (Francis *et al.*, 1996).

The sentence has sometimes been chosen as an important category, since it too is fairly unambiguously identified, but this can be problematic. For example, Hughes (1994), attempting to pin down the roles of different parts of speech and indeed of individual lexical items, started by plotting the typical positions within the sentence of prepositions, pronouns, etc. Each initially appeared to have its characteristic profile. However Hughes found that this profile was muddied by the fact that one sentence might consist of one, two or more clauses. He had to abandon the sentence as a category and examine instead the neighbouring word; by analysing a large quantity of text he was able to cluster words semantically on this basis alone.

Stubbs (1996) has worked on key words also using a word-and-collocation-span model. His results are akin to those of Williams, though based on processing quantities of text which Williams could not have dreamed of examining a mere two decades ago. For Stubbs now, as for Williams in the 1970s, a key word is a pointer to social attitudes.

> The study of recurrent wordings is therefore of central importance in the study of language and ideology, and can provide empirical evidence of how the culture is expressed in lexical patterns. (Stubbs, 1996: p. 169).

The claim being made here is that the starting-point to a considerable degree determines the Corpus Linguistics tools; these determine the kinds of patterns which can be found. By analysing words within a narrow immediate context of a few words to left and right (chiefly to the left), a very large number of valuable observations have been made by those authors already referred to, about the English language in general, and its characteristic lexico-grammatical patterns.

A quite different perspective arises if one starts from the category *text*. Witness the work of Hoey, who has shown (1991) that it is possible to view lexical links, which make sentence bonds, as indicators of text structuring and even as signals of potential abridgement, or (1995) that lexical patterns can reveal textual (as opposed to grammatical) patterns. Berber Sardinha (1995), likewise, using the key words procedure to be described herein, has shown that lexico-text linkage can operate between clusters of texts on the same topic even if written by different journalists.

In such a perspective, one may do most of one's work within the English language, using the genre written journalism (largely because of the availability of large corpora of newspaper text), and one may require the category word for analytical purposes; but the aim is first and foremost to characterise texts, and from there, in the present paper, to develop means for drawing inferences regarding the culture these texts spring from. The text as a category is easily identified; tools for examining texts will give rise to a different class of insights than those devised for studying nine-word contexts, the language or a genre. The key words identified here will show a rather different kind of association-linkage from that of Stubbs: they will go some way towards accessing cultural schema availability as provoked by a key word.

THE STUDY

This paper reports on a set of procedures for studying schema patterns as reflected in texts. To illustrate the procedures, nearly 5000 texts were processed. After briefly describing the context of the study, we shall examine the various procedures, which derive from and build on one another. Thus the word listing procedure enables key word processing. This in turn allows for the creation of a 'keywords database' which will reveal 'key-key-words' (sic), and allow us to identify 'associates' which, as we shall see, can be clustered together in 'clumps'.

The procedures are those built into *WordSmith Tools* (Scott, 1996) and further developments to this suite of lexical analysis tools which I am working on, and which I hope to make available to users.

The texts used for this study were features stories from 1992-4. These were chosen instead of sports, home news, foreign news, etc., because they discuss a wide variety of topics, not always immediately linked to an event of the day's paper, and because these topics might be supposed to provide some insights into the current preoccupations of *Guardian* journalists and *Guardian* readers broadly representative of the political centre of the educated middle classes in Britain.

THE NOTION OF A KEY WORD

A key word may be defined as a *word which occurs with unusual frequency in a given text.* This does not mean high frequency but unusual[5] frequency, by comparison with a reference corpus of some kind. In the present study, the reference corpus consists of just over 70 million words of Guardian text from 1992–4.

The procedure for identifying a key word has several stages. First, a word list is computed, containing all the different types in the reference corpus and the frequencies of each. In the present case, the most frequent word was *the*, which came over 4.5 million times in the 70 million words, i.e. 6.38% of the running words were *the*. Next, the same sort of word list is computed for the text whose key words one wishes to find. In the present case, there were a lot of separate texts (all 4672 features stories from the same Guardian corpus[6]) but the procedure may be carried out with just one.

Third, each word in the individual word list is compared with the reference corpus word list. The item *the*, for example, usually takes up about 6% of the words in this type of corpus. If the percentage is similar, the item may be ignored. Where there is a great disparity in frequency, however, it is possible to identify an item as key. The actual calculation of "keyness" is done using the chi-square statistic,[7] but the important point to grasp here is that the notion underlying it is one of outstandingness. In other words, if a word occurs outstandingly frequently in our text, it will be key. Finally, when all potentially key items have been identified, they are ordered in terms of their relative keyness.

This is best understood by reference to an example. Here is the beginning section of one of the features stories:

> ***A woman's place is in Rome***
> A new group starts campaigning next month for the ordination of women in the Catholic Church. Twenty centuries of history are stacked against any change
>
> PAULINUS BARNES
>
> THE vote of the General Synod to ordain women priests was seen by some Roman Catholics as absolute proof the Church of England had finally lost its theological marbles. Some, but not all. For other Catholics, the vote was a cause for celebration—and the impetus to set up a campaign for their own women priests. They know they will have more than John Selwyn Gummer to contend with: 2,000 years of history are stacked against them. Not to mention a church hierarchy that is implacably opposed, almost to a man.
>
> The group, Catholic Women's Ordination, will be formally launched next month with a "vigil of mourning for women's lost gifts" outside London's Westminster Cathedral. Nicky Arthy will be there: for four years she was the Roman Catholic chaplain at Goldsmith's College and St George's Hospital Medical School in London. It was a happy and fulfilling time. But for Mass, the liturgical centrepiece of the community's week, Arthy had to invite a priest in.

And here are the first few (of 51) key words of the whole text as identified by *WordSmith Tools*, ordered in terms of their keyness:

ARTHY	3	(0.3%)	5	
MINISTERED	2	(0.2%)	14	
APOSTLES	3	(0.3%)	86	
CARPENTER	5	(0.5%)	324	(0.000005%)
CHRIST	7	(0.6%)	983	(0.000014%)
ORDAIN	2	(0.2%)	88	
ORDINATION	6	(0.6%)	837	(0.000012%)
CHURCH	19	(1.8%)	9769	(0.000138%)
SYNOD	3	(0.3%)	331	(0.000005%)
PRIESTS	6	(0.6%)	1383	(0.000019%)
CHAPLAIN	2	(0.2%)	200	(0.000003%)
LAPSED	2	(0.2%)	206	(0.000003%)
WOMEN	26	(2.4%)	34140	(0.000481%)
CATHOLIC	9	(0.8%)	4280	(0.000060%)
THEOLOGICAL	2	(0.2%)	280	(0.000004%)
JESUS	4	(0.4%)	1165	(0.000016%)

The first column of numbers shows the frequency of each item in the Church story; and this is followed by its percentage. The last two columns show the frequency in the 70 million word corpus, followed by their minuscule percentages. Thus (Nicky) Arthy is mentioned three times in the text and Arthy occurs five times in the whole corpus. Jesus comes only four times in the Church text and many more times in the whole corpus, but the percentages show that Jesus is proportionally much more frequent here. In effect, the reasoning is that Jesus occurs here many more times than one would predict on the basis of the 70 million word corpus.[8]

There are two thresholds to be set when computing key words. One concerns the chi-square cut-off point, which is set by establishing a minimum significance. In this case it was 0.000001. The other is a minimum frequency requirement: in this study, this was set at 2, which means that any word which occurred only once, however outstanding it might otherwise seem, would be ignored.[9]

KEY-KEY-WORDS

Having established the notion of a key word, a next step for current purposes was to create a database. By processing texts in large batches, it was a simple matter to create nearly five thousand word lists, one for each of the features stories in the sample. These were similarly batch-processed to provide key word files. And by sorting out the key words thus obtained, it was possible to identify "key-key-words": *words which are key in a large number of texts of a given type.*

A sample of the most key-key-words from the 4672 files is reproduced here.

HER	333	7.1%
I	329	7.0%
SHE	296	6.3%
HE	295	6.3%
LABOUR	281	6.0%
HIS	275	5.9%
WOMEN	258	5.5%
ELECTION	229	4.9%
DIED	202	4.3%
FILM	201	4.3%
MY	199	4.3%
PARTY	199	4.3%
MUSIC	195	4.2%
BOOK	190	4.1%
SYSTEM	189	4.0%
YOU	189	4.0%
WE	186	4.0%
EDUCATION	184	3.9%
WAR	178	3.8%
WAS	168	3.6%
TORY	168	3.6%
SAYS	160	3.4%
OUR	158	3.4%
POLITICAL	155	3.3%
DR	154	3.3%
ME	154	3.3%
ART	153	3.3%
IT'S	151	3.2%
POLITICS	151	3.2%
LORD	149	3.2%
WRITING	144	3.1%

The whole list contains 30 408 items, representing 193 213 occurrences *in toto*. To illustrate from the last word in the table, *writing*, this word is key (as defined above) in 144 files, i.e. 3.1% of the 4672 text files. (The table does not tell us how many times *writing* cropped up in each of these files.)

It is noticeable that there are many interpersonal items (e.g. *he*, *my*, *you*, *we*) in this listing. This probably reflects genre characteristics of features stories within journalism. Features tend to be more personal than City news stories. The other key-key-words give an initial indication of the kinds of topics which are of interest to features readers: politics, the Arts, the role of women, education, conflict.

THE ASSOCIATES OF A KEY WORD

The next stage of processing concerned identification of "associates". These are *words found to be key in the same texts as a given key key word.* To identify the associates of, say *women*, the Associates procedure went through the 258 files in which this item was key,

seeking out all the other key words in these key word files. After sorting by frequency, the results were as shown below.

women's	78
she	69
men	67
her	64
woman	60
male	50
female	49
I	34
sex	34
feminist	26
sexual	26
children	24
feminism	24
health	23
tel	22
contact	21
it's	21
mothers	21
my	21
says	21
child	20
baby	19
herself	19
mother	19
she's	19
work	19
equal	18
married	18
pregnancy	18
rights	18
book	17

The numbers are numbers of files: *book* came as a key word in 17 of the 258 files in which *women* was also a key word. Again, the table does not tell us how many times *book* appeared in the 17 files, merely that it was present.[10]

What the table does suggest is that women are perceived and described in features stories as being associated with a number of different roles and issues. Many of the items at the head of this list are those traditionally associated with women, but there are some which show late 20th Century preoccupations with women's rights (*tel* relates to telephone contact numbers for campaigns) and women in work.

This notion of Associate is strikingly close to the early 1950s and 60s discussions of collocation. Firth proposed the term in 1951 in relation to a narrow contextual environment, and immediately saw its utility in stylistics. He also saw that collocations can be related: "... the association of synonyms, antonyms, contraries and complementary couples in one collocation" (1957: p. 199).

Sinclair, writing fifteen years later, reconsiders Firth's term and its narrowness of context. He points out that "the stretch of text is *not* given" but concludes that "the implied definition of the environment is that it is the extent of text which is relevant in the description of an item" (1966: p. 412). After considering ever-increasing contexts going up to about 40 words, he decides that there might be a pattern such as *post, letter, pillar-box* "but it would be difficult indeed to devise a workable procedure for recognising it"[11] (1966: p. 413). He concludes by rejecting "for the moment, the suggestion that degree of proximity within the chosen boundaries of collocation should be considered of primary importance" (1966: p. 414).

It is noteworthy that, while these discussions of collocation gave examples like *letter, pillar-box, post* (as opposed to *letter, to, is, the*), computer-aided work relying on collocation horizons has been essentially unable to capture these associative links because a collocational span of four words to left and right of a search word is often too narrow to reveal the presence of such items, while a broad span produces far too much unrelated noise. By working not with a collocational span but the text-based notion of key word association, i.e. co-keyness, this problem seems to be overcome. That is, co-associates are not the same as Firthian collocates, but they represent a level of lexical patterning which inherits from Firth's traditions.

GROUPING ASSOCIATES IN CLUMPS

Associate lists are quite revealing, but a further stage was felt necessary for the present study. Many key key words are very likely to play numerous different roles within a corpus of texts. If they are central to lots of texts, this might be because there is a burning issue of the moment (mad cow disease at the present time) or because, as in the case here, they are deemed important at any time in late 20th Century journalism. If women get pregnant and have babies but also write books and work, so *systems* play a role in local government, in business, in computing, in the legal apparatus and so on. What was needed was a way of re-grouping associates. Hence, a *clump* of associates is *a set of associates formed by co-occurrence in the same texts which gave rise to associates.*

To compute these, the associates are re-computed as before, but where 10% or more of the items in any given keywords file overlap with those in another, the two are merged. After this, any associates which occur once only are eliminated from further processing, and the sets are displayed. Subsequently it is possible to re-group again using any percentage of overlap as a criterion. Using this procedure a series of clumps was computed.

At the first stage, 89 clumps were identified. These were similar to the ones to be shown below, but appeared to need further clustering. For example, clump 32, consisting of *film, funny, men, women* overlapped with clump 72: *she, film, career, women.* Accordingly a 20% second stage was done, which reduced the clumps to 57. Space does not permit reproduction of all, but a sample is shown below (Fig. 1).

Each clump thus identified points clearly to a different schema in which women play a role. The co-members of the clump are identifiable as parts of the stereotype: pregnancy is

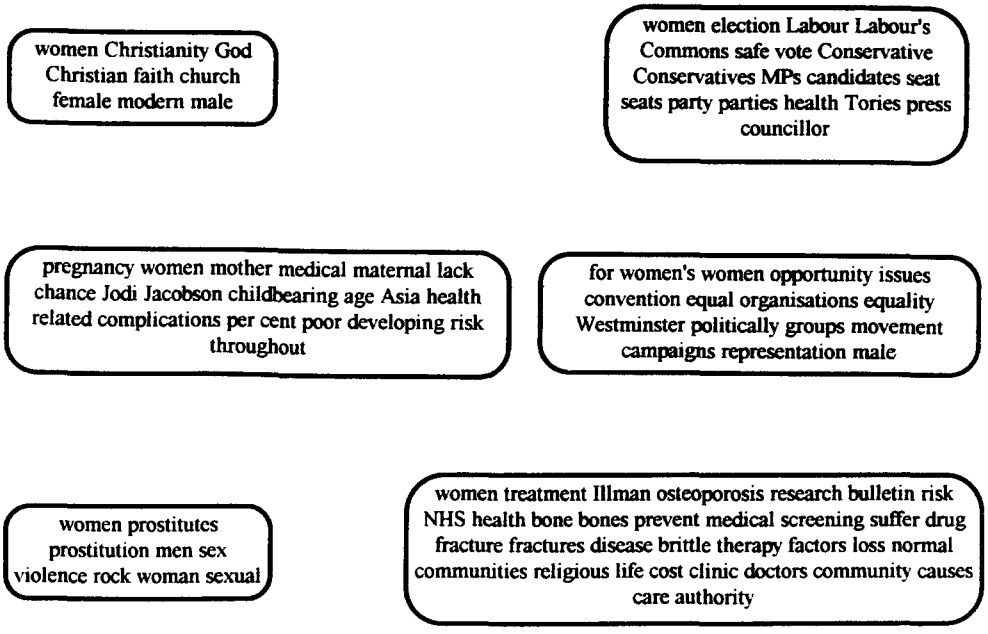

Fig. 1

not always at risk in the developing world, but the clump promotes an association which suggests it is. It is likely that the original articles, by dint of repetition of these issues, engender a similar effect in the uncritical reader.[12]

Further clumps related to issues such as: women in the media, abortion rights in Ireland, race, employment rights, HIV, sexuality, relationships, feminism, consumer affairs, domestic violence.

The procedure thus allows for differing percentages of overlap, which generate a greater or lesser number of clumps. There is no inherent desirability in reaching a small number of clumps. Originally, my intention was to use the procedure to disambiguate items such as *bank*, and there the aim would be to reach two clumps dealing respectively with rivers and finance. I reasoned that it was most unlikely for bank{river} to be a key word in any text where bank{finance} was also key. Put differently, this means it is unlikely for a key item to be ambiguous in one textual context.

However, it soon became clear that this procedure would help in the study not only of polysemy but also of stereotype and schemata. The clumps produced by higher or lower percentage settings vary in delicacy in Halliday's sense: at the most delicate level we may find a clump derived from two texts only, while at a coarser level it may be formed from the key words of many. The procedure has to allow the user to vary the percentage setting according to the delicacy required. In the examples shown above, women = [+female], [+adult] throughout; there is no ambiguity. Nevertheless women participate in a large number (be it 89 or 57 or some other number) of issues and contexts, and at present I do

not see how the number can be delimited in advance or whether it is reasonable to try. Further research is clearly needed.

In a further study carried out using the same methods but where two-word word lists were made, clumps relating to *The British* were as follows (Fig. 2).

These suggest that *the British* (as opposed to Britain[13] related to nuclear weapons policy,[14] to the Monarchy, to security issues, to European relations, and British weather!

Both the one-word analysis and this second example, based on two-word clusters, are presented here simply to illustrate procedures. It is important to stress, though, that the two-word cluster analysis is still quite superficial in comparison with the one-word analysis. Computing multi-word cluster word lists is extremely expensive in machine memory, since very few repeats are found. The percentage of hapax legomena (items occuring once only), already high at around 50% in single-word frequency lists, goes up dramatically with the number of words in a cluster. Computing cluster lists using virtual memory storage, i.e. on disk, soon becomes expensive in time. Here again, further work is needed.

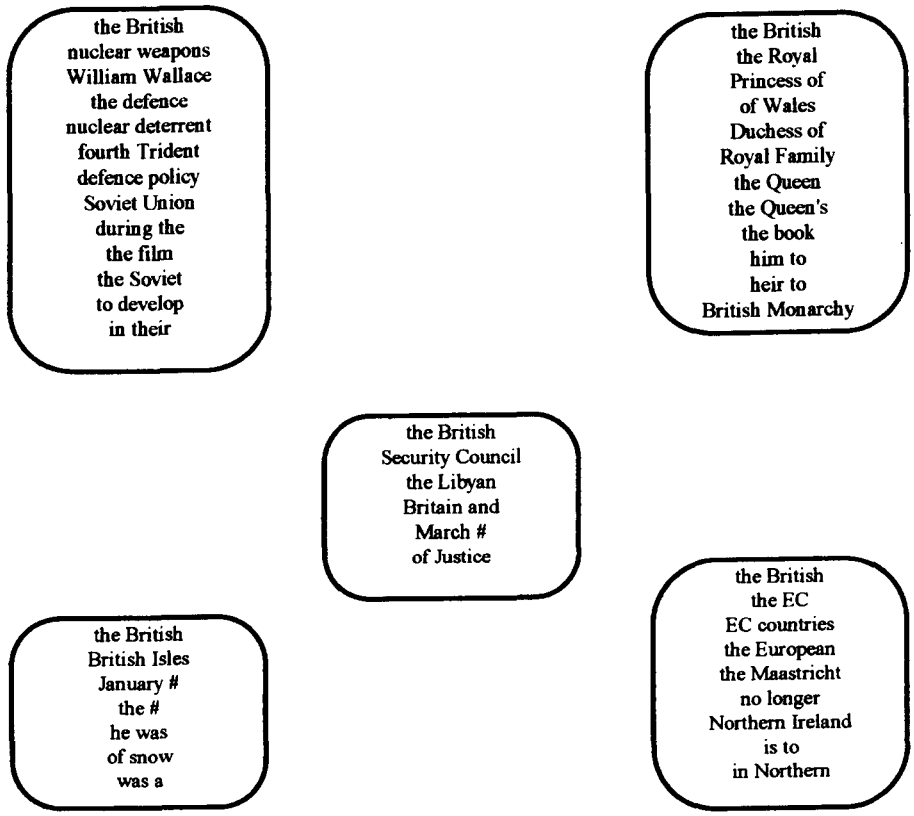

Fig. 2

CONCLUSIONS

The studies reported here describe procedures for tackling the relationship between writer schemata and text. By processing the key words in a large number of typical newspaper texts, it was found possible to identify clumps of associated key words, which appear to characterise current preoccupations, stereotypes and world content schemata.

The procedures for clumping are not yet refined; it is likely that methods discussed in Hughes (1994) will help, and certainly further research is needed with more examples of different corpora of texts belonging to differing text-types.

There are implications. Four come to mind immediately. First, if this procedure is found to be productive on other samples, it would provide guidance for text retrieval software design. Second, there are educational implications, starting with the identification of key words and clusters to be taught in EFL.[15] Third, key word clumps and their members will play a part in critical text analysis within Sociolinguistics, Applied Linguistics, Political Science, etc. And fourth, the notion of key words and associations between them has a role in literary criticism.

NOTES

[1]I am grateful to Mike Hoey and Tony Berber Sardinha for comments on an earlier draft of this paper.

[2]Cf. Richardson (1995) and Stubbs (1996) for penetrating discussion of Williams' central role and his methods in Cultural Studies.

[3]As a social phenomenon, stereotype has to do only with words but also with beliefs and concepts. The present paper does not claim to do more than gain access to some of its lexical manifestations, making no attempt to investigate belief systems behind them.

[4]As well as to the linguist when writing as opposed to analysing language. The fact that linguists have professionally had little or nothing to say about categories of importance to non-linguists, such as paragraph or chapter, may perhaps parallel the observation that non-linguists do not call on linguists for aid when in communication difficulties. The government may call on economists for help, but when a matter linguistic arises (such as what is implied by federalism in the EC) the OED is assumed to be the solution. It would be interesting to learn whether there are other fields in which both professionals and non-professionals use categories which the professionals do not investigate.

[5]Strictly speaking the frequency may be either unusually low or unusually high. The key words as described in this paper come from the unusually high ones. The KeyWords tool will also produce "negative-key" words, those whose frequency is unusually low. However, such items will not be accepted into a keywords database.

[6]All the features articles were used except for those of less than about 400 words

[7]Although the chi-square procedure has been criticised (Stubbs, 1995 and Dunning, 1993) insofar as lexical distributions within a corpus are not in any sense "normal", this procedure can be defended, e.g. by Rayson *et al.* (forthcoming) who used it to get at social differences in lexical usage with reference to British National Corpus spoken data. Technically there is no reason why one cannot use chi-square to relate the frequency of a word in one text to its frequency in a corpus, as long as one is considering the same word in each case. It all depends on what one is trying to *claim* on the basis of the chi-square procedure. The misgivings have to do with the skewed nature of types in a corpus and the very high incidence of singleton items. If there were 10 occurrences of "beetroot" in a 1000 word text on gardening, and also 10 occurrences of "beetroot" in a 1 000 000 word corpus of general texts, then that item would be 1000 times more frequent than expected on a chance basis and chi-square would be a reasonable way of saying that the difference is believable. If the occurrences were 1/1000 and 1/1 000 000, respectively, the same logic applies but the confidence in results differs, because 1/1 000 000 suggests the item is very rare and very rare items will *not* be spread around all possible corpora very very thinly (at a uniform rate of one per million words), but will crop up occasionally in relation to some sort of topicality or stylistic factor. In Rayson *et al.* (ibid) the items analysed occurred with quite high frequencies.

[8]Thus Arthy is the most "key" word, more outstanding than the others. Ordering of key words is done by computing the chi-square statistic and then sorting the list of words on this basis.

[9]In the KeyWords procedure one would not be considering a word which occurred less than twice in the text in question. However one might well have a word with a low total number of occurrences in text + corpus (e.g. three occurrences in the text and none in the corpus, giving a combined frequency of 3 in 1 001 000 words). The total would not meet the chi-square requirement of expected value $\geqslant 5$ in each cell. The item in this case would be mentioned in the keyword listing, but would not get the p value specified. In other words, the 3/1 001 000 suggests the item here *is* key (else why is it mentioned three times in a short text) but as it's very rare in the corpus we cannot prove it. The best justification, however, is that it works! I have done a lot of work now using various corpora in English and in Portuguese and have found it to be robust (results are quite similar even if the reference corpus is altered) as long as the reference corpus is fairly sizeable ($\geqslant 1$ million words, say).

[10]And as we know, it must have appeared at least twice per file to get into the key words listings.

[11]At this early date Sinclair recognised that intuition is "unreliable and extremely tentative. Our techniques are perforce based on the capabilities of machines..."

[12]Not exclusively the uncritical, of course. If all one ever hears about Liverpool is the Beatles, shootings, Jamie Bulger, football hooligans, one is liable simply not to know anything else about it. For British readers, ask yourself whether you ever hear anything about Brazil except destruction of the rainforests, street children, football, coffee and corruption!

[13]It is likely that the two-work cluster *the British* was most often part of a larger construction, e.g. the British policy, the British government's attitude, etc. The ♯ symbol stands for any number.

[14]The name William Wallace is borne by a present-day expert on European defence issues, not only the medieval Scottish leader. Unwitting lemmatisation rules supreme!

[15]This paper has concentrated on describing and illustrating a number of procedures; applications and implications for teaching have received short shift. The procedures do not provide a simple recipe for vocabulary work in L1 or L2, since vocabulary selection principles cannot simply be reduced either to frequency, to utility, to learnability or to text-importance. However, it should be clear from the data presented that these computational procedures do begin to relate frequency to text-importance, and thence potentially to generate many of the vocabulary items with which a generally key word can be associated. How they could become associated in a learner's mind is of course quite another story.

REFERENCES

Bartlett, F. C. (1932) *Remembering*, Cambridge University Press, Cambridge.

Berber Sardinha, A. (1995) *Intertextual Lexical Cohesion in Newspaper Reports*, paper presented at 40th International Linguistics Association Conference, Georgetown University, Washington.

Dunning, T. (1993) Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* **19**(1), 61–74.

Firth, J. R. (1951) Modes of Meaning, *Essays and Studies*, The English Association. Reprinted in *Papers in Linguistics 1934–51*, 1957, Oxford University Press, Oxford

Firth, J. R. (1966), *The Tongues of Men & Speech*, Oxford University Press, Oxford. *Speech* first published 1930.

Francis, G., Hunston, S. & Manning, E. (eds.) 1996, *Grammar Patterns 1: Verbs*. Harper Collins London.

Hoey, M. (1991) *Patterns of Lexis in Text*, Oxford University Press, Oxford.

Hoey, M. (1995) The Inseparability of Word, Grammar and Text, Inaugural Lecture, University of Liverpool.

Hughes, J. (1994) Automatically Acquiring a Classification of Words. PhD dissertation, University of Leeds.

Rayson, P., Leech, G. and Hodges, M. (forthcoming) Social differentiation in the use of English vocabulary: some analyses of the conversational component of the British national corpus, *International Journal of Corpus Linguistics*, Amsterdam: John Benjamins.

Richardson, K. (1995) Keywords revisited: the present as history. *Social Semiotics* **5**(Suppl. 1), 101–118.

Rumelhart, D. E. (1975) Notes on a schema for stories. In *Representation and Understanding: studies in cognitive science*, ed. D. G. Bobrow and A. Collins Academic Press, New York.

Scott, M. (1991) Demystifying the Jabberwocky: A research narrative. PhD dissertation, University of Lancaster.

Scott, M. (1996) *WordSmith Tools*, Oxford University Press, Oxford.

Sinclair, J. McH. (1966) Beginning the study of lexis. In *In Memory of J. R. Firth*, ed. C. E Bazell, J. C. Catford, M. A. K. Halliday and R. H. Robins. Oxford University Press, Oxford pp. 410–30.

Stubbs, M. (1995) Collocations and semantic profiles: on the cause of the trouble with quantitative studies. *Functions of Language* **2**(Suppl. 1), pp. 1–33.

Stubbs, M., 1996, *Text and Corpus Analysis*. Blackwell, Oxford.

Williams, R., (1976), *Keywords*, Fontana, London.