The Automatic Selection of Concordance Lines

Alex Collier

PhD 1999

The Automatic Selection of Concordance Lines

Alex Collier
University of Liverpool

ABSTRACT

This thesis presents the results of an experiment into the automatic selection of concordance lines from very large corpora.

Corpora now exist which are in excess of 100 million words in size, but the increase in size of corpora brings with it certain problems. These problems are discussed in the light of information obtained from professional corpus users and the continuing centrality of the concordance as the main means of interpreting the contents of the corpus is highlighted.

A possible means of overcoming the problems associated with the use of large corpora is presented. This solution is based upon software which was designed for the purposes of textual abridgement, this being carried out via an automatic analysis of lexico-cohesive bonds within the text. An analogy is drawn between conventional text and concordances; this analogy is then further explored by processing sets of concordance lines with the modified abridgement software.

In order to determine the success of the approach in identifying concordance lines which illustrate key features of the node word, an evaluation exercise is carried out, involving expert corpus users as respondents.

Contents

List of Figures

List of Tables

Acknowledgements

To My Family ... for putting up with me

To Mike Hoey ... for putting me up to this

Introduction

Introduction

This thesis concerns itself with the problems of using today's large-scale corpus resources, from which the primary output is still the keyword-in-context concordance. The increase in size of modern corpora has led to the situation where the corpus researcher is no longer able to retain an overview of the entire concordance for a growing proportion of the word forms contained in the corpus. Existing analytical tools alleviate this problem to some extent, but what is required is software which will present a condensed or 'abridged' version of the concordance, highlighting those lines that are most useful. The approach presented here draws upon work in the creation of automatic abridgements of conventional text and draws parallels between the concordance and other recognised text types, with a view to establishing the applicability of abridgement techniques to concordance data.

This thesis is structured in a manner which reflects the design of the underlying research, which on the whole is similar in form to a chemistry experiment in which the software is the apparatus and the linguistic data, the concordances, take the part of the chemicals processed and transformed using the apparatus. An initial description of the Apparatus and the chemicals thus leads into a discussion of the Method by which the chemicals are transformed, which in due course brings us to the presentation of the Results of this transformation. The latter stages of the writing-up of this experiment naturally include a discussion of the validity of the apparatus and method and present a series of Conclusions as to whether the goal of the experiment has been achieved.

The ultimate goal of the research described herein is to justify the processing a set of concordance lines using the software described. This is the most difficult part of the exercise and unfortunately one which has no equivalent in the chemistry experiment analogy, being largely dependent on individual judgements by human beings based on their linguistic intuitions.

The presentation of the research adheres largely to the structure suggested by the experimental nature of the work. The first few chapters cover the 'Apparatus', examining first the ingredients and then the design of the apparatus itself. Chapter 1 describes the form of concordance lines, mentioning briefly how they are created and outlining the rôle that they play in the use of online corpora. It examines the uses to which concordance lines can be put and the type of information that can be extracted from them. Chapter 2 presents a summary of the means by which large corpora can be exploited and introduces the problems that can be involved in this, based on the views of expert users of one of the large corpora. Chapters 3 and 4 introduce the theme of lexical cohesion and examine the feasiblity of using cohesion-based analytical methods on corpus data. Chapter 5 devotes itself to the description of software which has been developed to perform cohesion analysis automatically, firstly with the aim of performing automated abridgements of texts and latterly for the purpose of analysing concordance lines. In Chapters 6 and 7, the 'Method' of the experiement is introduced by means of a description of the parameters to the software, *cohort*, used in the course of my research and of the effects of combining them in various ways. Chapter 8 contains several sets of output from the system – the raw results of the experiment, to continue with that analogy – and it is not until Chapter 9 that an account is presented of the detailed evaluation process which was carried out on the results of the automatic system, correlating these with the intuitions of a group of experts and arriving at conclusions as to the optimal set of parameters for the selection of concordance lines fitting various criteria. It is this chapter which attempts to bridge the gap between the Results and the Conclusions, with the latter half of the chapter most strongly resembling the 'Conclusions' section familiar to chemistry scholars. In the final chapter, several further avenues of research are introduced, some closely related to the research described herein and others representing a considerable abstraction from the core principles of cohesion.

A Note on Terminology

One of the main concepts with which this thesis concerns itself is the *concordance line.* A set of such lines is generally referred to as a *concordance*, although for some corpus users "a concordance" would be a single line and a *concordance set*, or simply "some concordances", would be used to refer to a number of concordance lines. This thesis will adhere to the former convention, so that 'concordance' may always be understood to mean a set of concordance lines.

References to Other Work

This work combines research in several different areas of linguistics: concordances, collocation, lexical cohesion and abridgement, whilst also covering some of the computational aspects which have been involved. The use of separate literature reviews for each of these various strands would detract considerably from the readability and logical structuring of the thesis and so external references have been incorporated into the individual chapters which relate to the different themes.

Chapter 1

The Concordance

1. The Concordance

1.1. The Development of the Concordance

Concordances are not a recent development. As early as the thirteenth century, concordances to the Bible were being created in paper form. Cruden, in the Preface to the 9th Edition of his Bible concordance (Cruden 1828), tells us of one such concordance, undertaken by Hugo de S Charo with the assistance of 500 monks. The size of Hugo's army of amanuenses gives a fair impression of the scale of this undertaking and a brief glance at the extent of a modern biblical concordance such as (Even-Shoshan 1977), running to three volumes, only serves to reinforce this impression.

Hugo's completed work listed all the 'common' words found in the Bible in alphabetical order, giving book, chapter and verse references alongside a single line of context. The approach and format used for these early biblical concordances survives to this day, although few scholars can call upon so many assistants. The basic idea is that all the uses of a particular word are gathered together in one 'list' with a certain amount of context visible to one or both sides of the word and, usually, a reference to the source location. By this means, the concordance user is able to gain an insight into the importance or behaviour of, for example, a given biblical character. It was the intention of the creators of such concordances that they be consulted alongside the original text of the Bible, the references allowing the reader to refer back to the full text. In exactly the same way, the modern concordance, drawn from an online corpus of text, tells us about the behaviour and importance of a particular word or phrase and allows us a window of context onto the text which surrounds this key or 'node' word. As an interesting parallel, most corpus retrieval packages allow the user to 'refer back' to the full text of the corpus from which the concordance was built and see the contents of a particular line in an expanded context.

Quite apart from the scale of their enterprise, the creators of early concordances were faced with a difficult task. The medium in which they worked was comparatively

inflexible. The book which they wished to concordance was in printed or even written form; their only means of accessing it was to read it sequentially, unless they were able to call upon an intimate knowledge of its contents. Dealing with material the size of the Bible or the complete works of Shakespeare was a daunting task though, however good one's knowledge of the text. Given this 'input' to the process of concordancing, what of the process itself and the 'output'? The concordance's author would need to decide at the outset which words were to be included in the concordance (the 'targets') and in what format they were to be presented (whether to include left and right context, how much context, any references, such as chapter or page number, to the origin of the concordance line). The author would then have to locate manually all the instances of each of the words, transcribing each one before finally gathering together all the instances of each word and deciding how best to present them in the final concordance.

This approach is one which might be called a 'total concordance', since a set of concordance lines is produced for each of the target set of words. The concordance sets are then presented one after the other in some alphabetical or thematic order. An analogous approach is followed in some of the early uses of computerised concordance generation. In the latter half of the 1970s, McCarren (1977) produced a computer-generated concordance to Catullus using some of the first corpus-processing software ever. This used the now-familiar keyword in context or KWIC format, whereby the node word was presented in a central position with an equal amount of context on either side. The pioneering corpora such as LOB (Hofland & Johansson 1982) and Brown (Kucera & Francis 1967) were likewise sometimes made available as a 'total concordance'. This means that a concordance was generated for each word in the corpus and that all these concordances were then printed out or stored on magnetic tape for further computational processing. It should be borne in mind that computational resources were scarce and that by pre-processing the corpus in this way, no special computationally-intensive software was required by the corpus user; all the intensive work had been done in creating the

concordances. The concordance could then be examined simply in its printed form or, in its electronic version, by the use of rudimentary text display and search software. This approach persisted until the 1980s. The original Birmingham Collection of English Texts (BCET), on which the early Cobuild publications (Sinclair 1987a, 1990) were based and at that time the biggest corpus project, was also pre-processed in this fashion, with the resulting concordance sets being stored both on micro-fiche and in hardcopy form.

The implications which this kind of approach has for the form of the concordance are substantial. The effect of this process is to render the concordance immutable, almost as fixed as the paper concordances of the sixteenth century, more so in the case of the BCET, which ultimately contained twenty million running words. No wonder, then, that a micro-fiche version was created. The computational effort of creating such output was immense, given the resources available at the time. Clear (1987) tells us how the total concordance was created in batches, one letter of the alphabet at a time, on a mainframe computer, using up nearly its entire capacity during the weekend periods when there were few other users. The task was complicated by the additional job of sorting each word's concordance according to its right-hand context, so that, for example, lines for 'apple' where 'apple' occurred to the left of 'cart' came before those where it occurred adjacent to 'pie'. The result, however, as stated previously, was then as fixed as its thirteenth century manuscript predecessors. It is perhaps thanks to this fixedness of form that the concordance has endured. Even with access to modern sophisticated corpus searching software packages (WordSmith Tools (PC Windows), HUM (Unix), TACT (DOS), SARA (PC Windows/ Unix), Free Text Browser (Mac) or MonoConc (PC) to name but a few) which allow access to the full source text of a concordance line at the touch of a key, corpus researchers instantly understand the traditional format and are able to extract information from it. This has been a major factor in determining the input to the system described in this thesis. For reasons which will become apparent later, though, the concordance is not always as willing to render up its secrets as one might expect.

The last decade has seen a dramatic increase in the processing power of computers and a corresponding decrease in the price of processors and disk storage. New programming techniques based upon database technology have greatly streamlined the task of creating a concordance, be it for a single word, a phrase, or all the words in a corpus. This has given far greater flexibility to the form in which concordance data can be created and presented. It is now possible to pre-index a corpus so that concordance lines of any given length and context can be created almost instantly for any word or combination of words. Markers can be added to the lines to indicate the individual text of the corpus from which they came or any other textual or meta-textual information which has been included in the corpus. Lines can be re-sorted instantly to examine recurrent patterns in the environment of the node word in question. With this range of facilities at their fingertips, what uses can corpus researchers make of concordance data?

1.2. Direct Uses of Concordance Lines

It was mentioned earlier that a concordance can tell us about the behaviour of the *node word, that is, the word which was used to generate the concordance and which in the modern KWIC format stands centrally in the concordance line, as can be seen in the following sample concordance, where the node word is 'frog'.*

```
 that we could have  thought she was a frog.  {C} Was it Dave who said that. I
ot bobbing up and down like a demented frog saying,  take it to the United Nat
mprecations, ''eye of newt, and toe of frog"", contemplation and silent prayer
ed and for a second I thought it was a frog. I'd never seen a fish like that b
 more times and it still looked like a frog, but it didn't have any legs. Then
termediate forms which could be called frog or toad with equal accuracy. Inste
The biggest anuran of all, the goliath frog from West Africa, is able to jump
in effect, a small parachute. When the frog leaps off the branch of a tree, th
losive, so surprising, that catching a frog can be a difficult business, wheth
nsect larvae. In Brazil, another small frog builds its own ponds on the margin
```

Figure 1.1: Sample Keyword-in-context Concordance

From a simple inspection of a concordance there are a number of things which can be learned. Assuming that we are dealing with all the available concordance lines, rather

than just a sample, then without even reading a single word of the concordance, it is possible to gain an overview of the frequency of the node word. Obviously, since every occurrence of the node word generates one line of the concordance, a word which occurs many times will occupy several pages. We therefore instantly know how much evidence is at our disposal. Following this through, we can gain an impression of how representative the concordance data might be; no corpus is all-encompassing (i.e. it cannot contain the entire linguistic output of the universe)†, thus if only a few concordance lines are present we have to take it on trust that this is a rare word, a fact which we must bear in mind when we try to make any generalisations about the behaviour of the node.

By closer inspection of the concordance, it is possible to draw conclusions about the behaviour of the node word on the basis of its near neighbours or 'collocates'. Firth (1957) tells us that it is this very set of collocates, 'the company which a word keeps', which defines its meaning. Since the concordance is generally sorted on one of the words in the environment of the node, instances of repeated collocates will tend to cluster together. Thus in a concordance of the node word 'zone', if the concordance is sorted on the word to the left of the node, patterns of behaviour can be observed as the recurrent neighbours cluster in adjacent lines:

```
he was almost at the edge of his drop zone he was momentarily unable to do an
than five miles from the nearest drop zone. During, who commanded a heavy mac
s garden and started towards his drop zone north of Ste. Mere-Eglise, he hear
ad landed on the east side of the drop zone. Between him and Varaville were no
he 12th Battalion, miles from his drop zone, the first sound of war was a moan
llivan set out to reconnoitre the drop zone. Within minutes he was hit by fire
t instead of landing in a lighted drop zone he was heading for the centre of a
```

In a case like this, where the node is primarily used as a noun (exclusively so, in this selection of lines), we have chosen to sort by the word to the left of the node in order to establish which words can be used to premodify 'zone'. It is easy to imagine that patterns

† Much has been written on the representativeness of corpus material and on attempts to create a 'balanced' corpus, that is, one which presents a comprehensive picture of the language it seeks to describe. See Renouf (1987) for an excellent account of the composition of BCET. The quest to create the ultimate corpus can, however, become rather circular, since any all-encompassing or 'universal' corpus would have to include itself as part of its universe. The trend has therefore been to approximate to the universal corpus by creating ever-larger corpora, the theory being that the larger the corpus, the greater the correspondence with the universe of which it forms a subset. As we shall see in later chapters, this approach brings with it certain disadvantages.

of collocation such as 'drop zone' in the above concordance lines leap out at the corpus user and that the collocational profile which ultimately defines the behaviour of a given node can thus be easily extracted from the corpus data. This would be a logical extension to the point mentioned earlier that the number of concordance lines is an important feature of the concordance: 'drop zone' occurs, as can be seen instantly from the concordance lines, seven times, which should indicate to us that 'drop' forms a significant part of the collocational profile of 'zone'.

Sadly, the exercise is not always this simple. Let us now choose some more lines for 'zone', again left-sorted, but this time where the word immediately to the left is 'exclusion':

```
76> group close to the total exclusion zone and closing on elements of our tas
nverging at speed. The total exclusion zone, however, was 'not relevant in thi
t begins by referring to the exclusion zone and ends by claiming it wasn't rel
thirty-six miles outside the exclusion zone we had publicly set, and asked how
```

In this set of lines, two of the four occurrences of 'exclusion zone' are further premodified by the word 'total'. In this instance they happen to occur adjacently, making the pattern easily identifiable. In order to guarantee that such pre-premodifiers would occur together, it would be necessary to modify the sorting algorithm for the corpus retrieval software so that it performs a *compound* sort, sorting first on one word to the left, then two to the left and so on. This is described by the shorthand *-1, -2*, meaning sort on the word one position away from the node, the minus sign signifying to the left, and then within this, sort on the word two positions from the node, again to the left. This operation is quite feasible technically, but the approach which it implies is based on the assumption that, in this instance, 'exclusion' is the key word of the premodifying group 'total exclusion'. Compare this case with the following set of lines, sampled from a concordance of 'wrong':

```
untry (whether the country is right or wrong), we make it easier for them to g
 the chance to decide what is right or wrong for their country. ... First the
own conviction as to what is right and wrong. If conscience lands him in jail,
efined sense of morality, of right and wrong. I was not surprised when it turn
```

```
ay, in Graham's case, is that right or wrong I wish he hadn't gone. However, p
 'Whatever else you conclude, right or wrong, don't make any mistake about Hal
nge their mind. The ideas of right and wrong that their parents taught them ha
are still too young to know right from wrong, will enter the land - the childr
 woman who had taught Ginny right from wrong, the woman who had repeatedly ass
```

These were sorted on the word two to the left of the node word 'wrong'. There were many more occurrences of the pattern 'right *X* wrong' (50 for 'right or wrong', 54 for 'right and wrong', 4 for 'right from wrong' and just one 'right the wrong'), but these few will serve to illustrate the problem. As we would expect, the occurrences of 'right' cluster conveniently together, making this an easily identifiable pattern, but what of the word between 'right' and 'wrong'? There is obviously a sub-pattern in this slot, but it depends upon the presence of the word 'right' as the key word of some larger pattern and would require a *-2, -1* sort in order to identify it. Simple resorting or compound sorting is a laborious means of identifying such patterns, and since the corpus researcher may not always be aware of their existence, it is entirely possible that patterns may be overlooked altogether. Of course, the degree to which this occurs is difficult to measure, as one cannot ask a corpus analyst to count something which they have failed to recognise! There is, however, some evidence to be gleaned from the testimony of the users of large corpora encountered in the course of this study that the growth in corpus size is increasing the likelihood of this problem occurring.

1.3. Other Uses of Concordance Data

1.3.1. Collocates Revisited

In the previous section the concept of the collocate was introduced. It was shown that the set of collocates which belong to a given node word can be of help in the process of defining its meaning and that sorting and re-sorting of a concordance could be used to discover the collocates which regularly occurred in particular positions in the environment of the node. It is possible, however, to retrieve the collocational profile of a node

word automatically without first investigating its concordance. This is achieved by examining those words which occur within a pre-defined *span* or context of the node, noting any which re-occur with greater than random frequency. Various statistical measures, all based on the frequency of the node word, the size of the span and the corpus and the overall frequency of the collocate in the corpus as a whole, have been applied in order to identify the strength of association between a node word and its collocates.

One way in which this can be visualised is the 2x2 contingency table. For two words, X and Y, the table lists the four (hence 2x2) possible ways in which the occurrences of X and Y can 'overlap':

	Y occurs	Y does not occur
X occurs	A	B
X does not occur	C	D

Figure 1.2

2x2 Contingency Table

In this table, A represents the number of times X and Y co-occur, while the total numner of words in the corpus is given by summing A, B, C & D. The figures B and D show how often X and Y respectively occur independently of each other. Once the co-occurrence of X and Y are expressed in this way, they can be manipulated using measures such as Chi-square or log-likelihood. See Daille (1995) for a discussion of the relative merits of these.

Another often-used measure of significance is the Z Score (Butler 1985). It is based on comparing the number of times words X and Y co-occur (the *observed* co-occurrence) with the number of times that they ought to co-occur, given the frequencies of X and Y, the size of the corpus, and the amount of context around X and Y that is examined (the *expected* co-occurrence). This Z Score is calculated as a ratio of observed:expected.

Where the observed co-occurrence is more than about twice the expected, X and Y are deemed to co-occur significantly.

The Mutual Information (MI) Score (Church and Hanks 1989) and T Score (Church et al 1990) have also been used in identifying strongly associated collocates, although neither of these provides a measure of significance, but rather seeks to rank the collocates of a word in order of similarity (or dissimilarity in the case of the T Score).

Any of these approaches can be exploited to derive a list of words which regularly occur in the environment of a node word. The measure used in this study is the Z Score and the discussion which follows will use the term 'significant collocate' to mean one which has been selected on the basis of its significant co-occurrence with its node word as measured using the Z Score.

It is stressing here that the way in which cohesion analysis works is substantially different from any measures of collocation based on strength of association or significance. Such tests 'reward' what they regard to be unusual, in that the co-occurrence of X and Y more frequently than would be expected by chance is flagged in some way (high significance, strong association). To do this, these statistical measures have to make some recourse to outside data – generally the corpus (independent) frequencies of X and Y. The cohesion-based algorithm used by *cohort*, on the other hand, *simply measures what is there.* It looks for the presence of one or more 'collocates' in a concordance line and attempts to find other lines where those collocates are also present. In contrast to some of the measures outlined here, it does not impose an arbitrary cut-off based on an ill-fitting statistical model (see Drawbacks of Using Collocates, below). The identification of statistically significant collocates is not, therefore, a possible alternative to cohesion, but it will be used later on in helping to evaluate the effectiveness of cohesive analysis. From time to time in the thesis, reference will be made to 'collocates' that *cohort* has identified. This must not be confused with the significant/strong collocates highlighted by the statistical measures mentioned above. When using cohesion analysis to look at concordance lines, those

words which are identified on the basis that they are taking part in the formation of links between the lines of the concordance are referred to as collocates. The process of link formation is described in full in Chapters 3 and 4.

A list of the significant collocates of a node word can give corpus researchers clues as to which features of the concordance line they should be looking for in order to identify those lines which represent the best examples of the typical collocational behaviour of that node word. Suppose, for example, that the list of significant collocates for the node word 'zone' contained the word 'outside'. It would then be advisable for the corpus researchers to examine concordance lines for 'zone' in which 'outside' occurs in order to gain a fuller picture of the behaviour of the node 'zone'.

Several factors influence the contents of the collocate list, most notably the span and the significance level, and some care needs to be exercised in selecting these. If a large span is chosen, then the number of collocates identified will be greater, perhaps yielding a list which is unwieldy or diluted by uninteresting collocates. If too small a span is employed, then important collocational information may be missed. The optimal size of span is a topic of some debate and it is a topic to which we shall return later.

Let us now instead move on to the significance level. This is simply a threshold, based on established statistical models, which defines how high a Z Score for a given collocate must be in order to prove that the collocate is present in the environment of the node more frequently than would be attributable to chance alone. The score is based on determining the number of times a collocate occurs independently of the node word and extrapolating from this the number of times that it *ought* to occur with the node word. This figure is then compared with the frequency of co-occurence of the collocate and the node to arrive at a statistical measure of significance.

1.3.2. Drawbacks of Using Collocates

We have seen how a list of significant collocates, derived from the set of words within a closely-defined environment of a node word, can be of some assistance in identifying and typifying the recurrent features of a concordance. Such a list does not, however, tell the whole story.

Since the collocate list is an amalgamation of information based on all the positions (relative to the node word) inside the specified span, the precise location of each collocate is lost. Unfortunately, this cannot be overcome simply by using first a span of one word either side, then two words and so on, as we soon find ourselves in the same predicament that we encountered in the sorting examples above, where patterns of collocation can span several positions to either side of the node. Since a span of one has, by definition, no knowledge of the contents of a span of two, any inter-relationship between the columns is thereby lost. There are a number of other definitions of collocation which can take more information into account, but these do not tend to be applied directly to concordance data. Schütze and Pedersen (1993), for example, employ a much larger amount of context – anything up to 40 words – which could not feasibly be provided by a standard KWIC concordance and WordSmith offers the facility to analyse collocates up to 25 words either side of the node. The problem of the inter-relationships between the collocates is addressed by Brown et al (1992), who record the individual sequences of collocates by means of an *n*-gram model. While this goes some way, it still relies on a *n* being sufficiently large to encompass all the patterns present in the node word's context and may therefore not be successful where, for instance, a pattern straddles the node word. For large values of *n*, furthermore, the complexity of the approach increases significantly, resulting in the 'large parameter space' problem, noted by Stolcke & Segal:

> An *n*-gram grammar is a set of probabilities $P(w_n | w_1 \; w_2 \; ... \; w_{n-1})$, giving the probability that w_n follows a word string $w_1 \; w_2 \; ... \; w_{n-1}$, for each possible combination of the w's in the vocabulary of the language. So for a 5000 word vocabulary, a bigram grammar would have approximately 5000 × 5000 = 25,000,000 free parameters, and a trigram grammar would have

> 125,000,000,000. This is what we mean when we say *n*-gram grammars have many parameters. (Stolcke & Segal 1994, p 1)

As we have just seen, most methods of collocational analysis obscure the relationship between individual collocates within the context of the node word. Another problem arises because the link between the collocates and the original contexts is also lost. Let us assume for a moment that the collocate list informs us that 'war' is a significant collocate of 'zone'. Making use of this information is not simply a question of extracting from the corpus all concordance lines for 'zone' which contain 'war'. The fact that 'war' is an interesting collocate of 'zone' in the phrase 'war zone', does not exclude the possibility that there exist other lines where 'zone' and 'war' co-occur, but this may be as part of some other phrase or even in a phrase which is unique to a particular line. This is exemplified in the next figure, which lists the concordance lines for 'zone' which contain 'war' as a collocate.

```
 1  Arab Republic of the Red Sea as a war zone, and the closure of the Straits of
 2  an operation he'd just been on in War Zone C, above Cu Chi. "There were a lot
 3  only neutral countries within the War Zone were Sweden and Finland. The Swede
 4  that the Western Approaches was a War Zone, into which shipping of any kind e
 5 . include 50,000 refugees from the war zone Be- tween Ethiopia and Somalia, an
 6 ant ships approaching the declared War Zone to turn back. The next was to cons
 7 compliance. The declaration of the War Zone was not regarded by the Soviet Uni
 8 e and flew missions, mostly around War Zone C, along the Cambodian border, and
 9 ern Ireland is another ideological war zone she prefers to skirt around; yet 1
10 he 12th Battalion, miles from his drop zone, the first sound of war was a moan
11 of 13,000 people just north of the war zone, the hospital is taking a stead
12 ol prior to the declaration of the War Zone. Third, two diversionary fast conv
13 plements of war, and to create a peace zone in the Indian Ocean. But for every
14 thin the West Nile to  escape  the war zone.  Only one hospital,  at Angal,  i
```

Figure 1.3: Concordance Lines for 'zone' which contain 'war'

The problem mentioned previously is demonstrated in Figure 1.3 by lines 10 and 13, where 'war' is not part of any pattern which might be said to be collocationally related to the node word. Such lines serve to dilute the subset of authentic lines which contain the features which we are trying to identify. What is required is a method of analysis which only operates on lines which are known to contain a repeated feature, yet those features can only be identified by calculating their significance as collocates, which has to carried

out on the entire concordance, which means that the important information has already been lost because of the amalgamation of all collocate positions and the loss of the relationship to their original context. This approach is therefore circular and as such unworkable.

The actual means of creating the collocate list are also not beyond criticism. All statistical measures of significance rely upon comparing observed behaviour (the fact that words *a* and *b* co-occur) with some expected pattern of behaviour, generally based upon a model of normal distribution. Unfortunately, collocation, and language in general, do not adhere to any such model; thus any attempt to ascribe to them normal behaviour can only ever be an approximation to their true nature and the need for human interpretation will generally be required. This is echoed by Stubbs (1995: 48):

> We always start with intuitions about what is interesting to study, and intuition re-enters, in designing procedures and in interpreting findings.

1.3.3. Columns

Another means by which concordance data can be used to access the collocational profile of a node word is to build up a picture of the words which surround it on a column-by-column basis. To go back to 'zone' for a moment, this would give us something like this:

Node

|

10 in	27 the	63 the	13 the	18 the	31 the	15 the	13 of
7 to	27 in	34 a	12 war	17 of	8 a	4 of	11 the
7 and	18 of	11 in	9 a	16 and	5 to	4 in	11 and
6 the	12 a	6 this	7 free	8 in	4 was	4 and	5 up
6 of	6 to	5 an	7 drop	6 where	4 of	3 zone	4 is
5 was	6 into	4 that	5 nuclear-free	5 is	3 they	3 would	4 a
4 a	4 on	3 total	5 landing	4 was	3 on	3 to	3 not
3 it	4 from	3 his	5 exclusion	4 to	3 is	3 not	3 had
3 is	4 and	2 within	5 danger	4 for	3 i	2 wells	3 be
3 east	3 zurich	2 u	3 three-mile	4 between	3 all	2 that	2 which
2 within	3 zone	2 to	3 this	4 a	2 which	2 ste	2 to
2 those	3 was	2 its	3 free-fire	3 or	2 what	2 regarded	2 on
2 they	3 through	2 airport	3 erotic	3 on	2 that	2 on	2 mere-eglise
2 separating	3 outside		3 erogenous	3 he	2 refroze	2 mouth	2 line

2 poured	3 now	3 enterprise	2 were	2 one	2 miles	2 in
2 out	2 were	3 combat	2 we	2 not	2 is	2 he
2 on	2 is	3 buffer	2 they	2 no	2 heat	2 have
2 my	2 inside	3 british	2 then	2 lava	2 had	2 genitals
2 miles	2 create	2 tropic	2 such	2 into	2 four	2 by
2 found	2 beyond	2 temperate	2 so	2 interval	2 blakiston	2 at
2 declaration	2 becomes	2 taboo	2 she	2 had	2 a	2 army
2 countries		2 splash	2 named	2 for		
2 at		2 soviet	2 molten	2 as		
2 as		2 smokeless	2 into			
2 area		2 rift	2 i			
		2 private	2 c			
		2 one	2 but			
		2 n	2 at			
		2 impact	2 as			
		2 dry	2 almost			
		2 demilitarised				

Here, each position relative to the node word is represented by one of the eight columns, the first four columns standing for left-hand collocates and the last four for right-hand ones. This format gives a very clear picture of the most frequent collocates in each of the positions, with each one being ranked in descending order of frequency. As such it provides a fuller picture than the simple collocate list, which lacks positional information.

The disadvantage of this type of output is that, as was demonstrated with the collocate list, the relationship across the columns is lost. Thus, although it is obvious that 'war' is one of the most frequently used premodifiers of 'zone', there is no way of discovering (short of going back to the raw corpus data) whether it is '*the* war zone' or '*a* war zone' which is more typical.

1.4. A Note on Tagging

No work on part-of-speech (POS) tagged concordances has been carried out in the course of this study. There are several reasons for this.

Firstly, there are nowadays several sources of POS-tagged corpus data. You may buy your corpus pre-cooked, as in the case of the British National Corpus, or you may prepare it yourself, combining one of several publicly available POS taggers (qv Chapter 2) with

your corpus ingredient of choice (see Section 2.4 for suggestions). When this study was begun, the situation was somewhat different – little (reliably) tagged corpus data was available and so no attempt was made to integrate tagging information into the system.

Secondly, the aim of this study is to examine how the change in size of corpora has affected their usability. Since most of the respondents in the user survey originally used an untagged corpus (in the days when it was only 20 million words in size), the simplest and fairest comparison to make was with untagged data drawn from a much larger corpus. Of course, this raises the question of whether a large tagged corpus is easier to use than a large untagged one, but this would be a difficult avenue of research to follow, since who, having tagged their large corpus, would set about ignoring the tags?

Thirdly, no use of POS tagging was ever made in the automatic abridgement system on which *cohort* is based. Since one aim of this study is to compare the handling of text and the handling of concordances using similar techniques (i.e. cohesion analysis), the additional complication added by the introduction of POS tagging as a further parameter to the system is best avoided. If it were to be introduced, one would certainly want to compare the results obtained when POS was used as part of the analysis with those achieved when it was not. This would increase the number of possible parameter combinations (see Chapters 6 & 7) even further. In addition, since this study is to a large degree an evaluation of the *cohort* system, the addition of POS information would require the inclusion of a validation process of the tagging software and the tagset which it implements. This would go beyond the scope of the current study.

Chapter 2

Issues of Using Large Corpora

2. Issues of Using Large Corpora

2.1. Introduction

In this chapter we shall discuss the growing use of large corpora and examine the reasons why they have grown in scale and popularity in recent years. My interest in this area started while I was engaged in the creation of a large textual database for the ACRONYM Project (Renouf 1996). This corpus database was mainly required to underpin a database of collocational relationships, but was also, to all intents and purposes, a corpus, and was accessible using the corpus facilities outlined in the previous chapter. The amount of text in this database exceeded 400 milion words and I had noticed that the concordance line output from it was often unwieldy, since so many lines were produced by queries involving frequently occurring items. To ascertain whether this was a common problem, I sought the opinions of other users of large corpora.

What is presented in the sections which follow includes an example of a large-scale corpus system and a synopsis of the opinions of users of this corpus relating to the issues surrounding its continued growth. The effectiveness of some commonly-used tools will also be discussed in relation to the use of very large corpora. Finally, a concrete example of the implications of expanding corpora still further will be shown.

2.2. Increasing Corpus Size

The last three decades have witnessed a growing use of large collections of electronic text (corpora):

> Thirty years ago when this research started it was considered impossible to process texts of several million words in length. Twenty years ago it was considered marginally possible but lunatic. Ten years ago it was considered quite possible but still lunatic. Today it is very popular. (Sinclair 1991, p. 1)

and while growing in popularity, the corpora themselves are becoming larger:

> The one-million word corpus, which was long a standard size, has already been replaced by much larger corpora (witness, for example, the British National Corpus), dynamic monitor corpora, online textual databases, etc., and this trend is bound to continue. (Svartvik 1996, p. 9)

As the popularity of a corpus-based approach to language study has grown, corpora have become larger and larger, assisted by the increasingly powerful computing resources which have become available. While huge corpora are not necessary for many types of linguistic research, the current opinion among corpus builders, as evidenced by examples which will be introduced later in this chapter, would seem to be that the bigger the corpus one can make, the better the results obtained from it will be. The latter part of this chapter will, however, put forward an argument which contradicts this opinion to some extent, showing that current methods of accessing corpus data will ultimately fail to deliver the hoped-for improvement in results.

2.3. Why Increase Corpus Size?

The creation of any corpus is a considerable undertaking and no corpus exists for its own sake, but is rather built and maintained with a some purpose in mind, even if that purpose is to be as generally useful to as many users as possible, as in the case of the British National Corpus (Burnard 1995). Many corpora are still created today which are of a comparable size to the early (sub-million-word) corpora. This may be because the 'text' of the corpus is difficult to obtain, as is ofen the case with spoken corpora. The COLT corpus (Stenström & Haslerud 1995) of 'teen-speak' is one example of this, with a size of half a million words, transcribed from 50 hours of audio recordings. It may also be the case that a corpus is highly focussed on a particular, rare linguistic feature, as is the case with Marco's 36,442-word corpus of rhetorical functions (Marco 1999). Constructing a huge collection of text is in cases such as these either impossible or unnecessary. At the other end of the spectrum, there are corpus users who need vast amounts of data, for reasons which will be made clear later in this chapter. Given the existence of a user group for ever larger corpora and bearing in mind the considerable resources which are brought

to bear in creating a large-scale corpus, it would be useful to examine some of the benefits of making corpora as large as possible.

2.3.1. Maximising Accuracy

One of the things which the use of corpus data allows us to do is to arrive at an accurate description of particular features of the language in a way that is less influenced by our own, possibly idiosyncratic, intuitions, in as far as accuracy can be defined as corresponding to the representation of language given by a particular corpus. Imagine that lexicographers are attempting to decide whether 'rely' + 'on' is a more likely phrasal verb than 'rely' + 'upon'. They would like to use the more frequent form as the headword of a dictionary entry, with the less frequently occurring form incorporated into the entry as an alternative usage, yet intuition cannot tell them for certain which occurs with the higher frequency. By examining the corpus, it will become instantly clear which is the more frequent, and the lexicographer can create the appropriate definition, safe in the knowledge that they have accurately reflected the commoner linguistic practice.

The corpus-based approach typified by the 'rely on/upon' example above is representative of the modern trend amongst writers of linguistic reference works. Whereas earlier works sought to prescribe 'correct' usage (Murray 1851), or document the language of the day to protect it from change (Johnson 1755), recent ones tend to be more data-driven, that is, they seek to describe the most typical features of the language (Sinclair 1987a). Corpus evidence provides a means of supporting this kind of analysis, since it avoids excessive reference to one's intuitions and can be updated to represent the most recent changes in the language.

Of course, any interpretation of the evidence from a corpus has to take into account the contents of that corpus. The points discussed here relate to corpora which are large enough and sufficiently balanced to provide a view of the language which is not skewed towards a particular style, genre or mode (spoken/written/written to be spoken etc). Any

corpus is only an approximation to linguistic reality, since no corpus may contain the whole language, especially where spoken data is concerned. The larger the corpus, it might be argued, the *closer* the approximation, as the corpus-subset becomes an ever-larger proportion of the universal set, language. It is for this reason that corpus builders are keen to create the largest corpus they possibly can. They can then claim to have the most accurate representation of real language.

By increasing the size of the corpus, it is also possible to achieve greater accuracy in any statistical processes which are applied to it, since a larger amount of data tends to yield more reliable results. This is applicable in cases such as the measurement of the statistical significance of collocation, or the comparison of the lexis used in different sub-corpora, where the more evidence one has, the more confidence one can have in the results obtained.

The goal of increased accuracy has always been a major factor in the *creation* of corpora, as well as in their exploitation. Early corpora were painstakingly checked so that any errors introduced when the text was typed or scanned from the printed page into the computer were corrected. Some corpora were manually tagged with part-of-speech, clause structure or prosodic markers, but such corpora have tended to be measured in terms of thousands of tokens. An example of this would be the London-Lund spoken corpus, in which the prosodic information was manually inserted, but which contained only 500,000 words. One might interpret this to mean that the size had to be limited because of the intensive task of adding the mark-up, but seen from the other perspective, it could be said that this was a justifiable means of extracting as much information as possible from the data, given its size; the corpus of the day represented a substantial investment of resources and it seems only natural to want to get the biggest possible return on that investment.

As the size of corpora began to increase, however, this perfectionist approach was no longer feasible, since hand-tagging or even comprehensive correction proved to be too

time-consuming. Even a corpus the size of BCET, some twenty million words, contained substantial numbers of errors, mainly introduced when the printed text was scanned into the computer via an optical character recognition system; thus there are dozens of occurrences of the word 'thc', caused by the OCR's failure to detect the cross-stroke in the 'e' of 'the'. Errors such as this can be fairly easily corrected automatically; there were, however, no resources available to manually check and edit every text in the corpus, and so many errors remained uncorrected. Despite the presence of these errors, however, BCET became a role model for future corpus production, since whatever errors it contained were outweighed by the sheer scale of the corpus as a whole. Given the large amount of corpus text which was free of error and the amount of human intervention that would have been required to correct the corpus, the effort of eliminating the errors could not be justified.

2.3.2. Maximising Completeness

Accuracy is not the only goal of corpus study – corpus users also seek to gain as *complete* a description of the language as possible. For as long as the corpus does not contain the entire language, a totally corpus-based description is not possible, but, as with the goal of increased accuracy, the larger the corpus gets, it could be argued, the closer it comes to the 'universal' corpus which was discussed in Chapter 1. As we shall see in the following section, where the working methods of professional corpus users are discussed, some intuition and interpretation will continue to be required in order to exploit the corpus to its best potential, however large it gets, since even the universal corpus, were it to exist, would simply *be* language and not a *description* of language. This relates to the point made earlier that no corpus can exist in its own right, but rather has to be created with some idea in mind as to how it can be exploited – it is otherwise just a collection of bytes on a disk drive.

The desire to create the most complete description of the language by means of the most complete corpus possible contrasts with the use of the pioneering corpora, the size of which limited their exploitation to the examination of frequently-occurring, largely grammatical, features. While even a modestly-sized corpus can tell us about many of the common grammatical and lexical features of the language, it will often prove unable to offer up any examples of the rarer phenomena. This shortcoming has fuelled the recent enthusiasm for ever-larger corpora – the bigger the corpus, the greater the chance that it will represent every possible feature of a language, always assuming that the corpus users are sufficiently well-equipped to find these features, which, as will be illustrated later in this chapter, is not necessarily the case.

The development of software (taggers and parsers) which will automatically assign part-of-speech or sentence-structure tags to corpus text has further increased the popularity of large corpora. Examples of these are CLAWS (Garside 1987), Brill (Brill 1992) and Helsinki (Voutilainen & Heikkilä 1994). None of these taggers and parsers can attain perfect accuracy (where perfection is equivalent to a manually-tagged/parsed text). Their usefulness comes to the fore in that they can be run over huge amounts of corpus data in a relatively short space of time. The effect of this is that although a few examples of a particular grammatical pattern may be missed because the software has assigned the wrong tags to it, the vast amount of correctly tagged data should more than compensate for the small number of incorrect ones. This is similar in principle to the argument which was mentioned in the discussion on accuracy: that a few errors are acceptable since they dwindle into insignificance in proportion to the overall size of the corpus, which is still able to deliver the hoped-for description of the language. The difference here, though, is that the added degree of inaccuracy is directly attributable to the shortcomings of the tagging or parsing system: if a grammatical structure is incorrectly analysed once, it is likely that this will occur at every one of its occurrences, while the correct analysis of previously unseen lexical items is dependent on the success of the tagger's morphological and

contextual rule matching. The mis-recognition by an OCR system of 'the' as 'thc', whilst certainly not random, is influenced by many more factors, such as the quality of the paper, age of the print and skill of the OCR operator. Naturally, the mis-analysis of a grammatical feature has important implications if one's motivation for increasing the size of the corpus is to identify new grammatical features, since if they do not correspond to the parser's idea of 'grammar', then they will probably be incorrectly analysed and therefore never recognized as new.

2.3.2.1. Corpora and Lexicography

Modern lexicographers and grammarians, using a corpus-based approach, no longer rely entirely on their own intuitions when attempting to decide what a word means, what patterns it forms with other words and in what circumstances it may be replaced by a synonym. Instead they call up some or all of the occurrences of the word in concordance form, drawn from a large body of text which is intended to represent real language. Intuition and evidence can interact in a variety of ways: at one extreme, it is possible that the researcher will have their intuitions confirmed by the information retrieved from the corpus; at the other, the corpus may not correspond at all to the user's conception of a particular linguistic feature. When this happens, the questions of corpus completeness, faith in the corpus and faith in one's own linguistic awareness arise, and users then have to exercise their judgement, or confer with others as to their impressions, until finally either their initial intuitions are deemed to be inaccurate or the corpus is judged to be unrepresentative, either through incompleteness or inaccuracy. Based upon the information passed on by the Cobuild corpus users and on personal experience, it would seem that it is possible to discover in the corpus some highly unexpected phenomenon, or to be surprised by the absence of an expected one, and that when this does happen it is not always attributable to the fact that one of the corpus text sources is somehow unusual in terms of its vocabulary usage or collocational profile, as might happen, for example, in texts drawn from a

technical or other specialised genre. We find this sentiment echoed by Leech who describes corpus examination thus:

> More detailed quantitative analyses (requiring large corpuses and the aid of computers) can be expected to produce results beyond the insight of a native speaker. (Leech 1966, p. 73)

The reference to the need for large corpora is an interesting one, being closely bound up with the central issue of this chapter. It begs the question, of course, of what was considered 'large' at the time Leech was writing.

With the corpus information at hand, the skill of the corpus lexicographer is put to work identifying the patterns exhibited in the concordance lines. As we saw in the previous chapter, this process can be automated to some extent; for example by providing a list of all the words which occur in the environment of the node word (the word to be defined), or by providing the facility to sort on various columns of words to the right or left of the node word.

When the corpus being used contains only ten or twenty million words, the evidence presented via the concordance lines is generally manageable enough, unless a very frequent word is being examined. Unfortunately, a corpus of such a size is coming to be regarded as too small for the satisfactory description of 'real language'. In addition to the issue of size, there is the question of currency to be addressed; the language is evolving constantly and any good dictionary or language teaching material must seek to reflect this. Since the dictionary is based on a corpus, the corpus too must be dynamic, kept up-to-date so that new words and usages can filter through into the publications which draw upon it as a source. While it could be argued that corpus builders should drop older material from the corpus as new material is added, so maintaining the corpus at roughly the same size and avoiding the problem of too much data, it has to be borne in mind that much effort was expended in putting together the corpus and simply to discard it might seem an imprudent waste of resources. This would also work against those analytical methods, such as the

statistical measures mentioned earlier, which operate more successfully as the corpus gets larger.

A corpus which is maintained at a (roughly) fixed size and contains only the most recently available material (e.g. the current years output of a daily newspaper) is known as a *snapshot* corpus. Of course, a number of snapshots may be created, each covering a different chronological period, in which case the overall amount of corpus data will still increase over time. Another means of limiting the size of a dynamic corpus is to use a *monitor* corpus approach (see Clear 1988), so that only interesting elements of the corpus are retained. An example might be to only store information (frequencies, collocates, word-class, context etc.) for those words which appeared for the first time in the most recent data. This is similar to the approach described in (Renouf 1993).

2.4. Existing Large Corpora

The perceived advantages of the corpus-based approach, coupled with the availability of large amounts of electronic text and the willingness of academic funding bodies to encourage this kind of research, have led us to the point where there are now corpus-building initiatives underway which are aiming to produce corpora containing in excess of 100 million words. The British National Corpus, advertised as being a balanced corpus of speech and writing, 100 million words in extent and fully tagged for parts-of-speech, was created as part of a UK government project within the Speech and Language Technology programme. It has recently been made available to the public on a set of CDROMs and occupies several gigabytes of disk space in its fully unpacked and indexed form. The software which accompanies it, while able to utilise the sophisticated SGML metalinguistic mark-up embedded in the corpus, provides only the simplest facilities for accessing the linguistic data. Other institutions have built up electronic text collections which exceed even this in scale. The ACL's Data Collection Initiative has already gathered hundreds of millions of words (although it makes no claims as to the balanced nature

of its holdings) and in the commercial sphere, HarperCollins Publishers have established a corpus system, the Bank of English (BoE), storing in excess of two hundred million words. Given sufficient disk space, it is reasonably simple to gather together vast amounts of textual data. The Research and Development Unit for English Studies at the University of Liverpool, for example, established online access to over 400 million words of newspaper text as part of the ACRONYM Project. All of this text was transferred from CDROM and frequency statistics from this corpus are included later in this chapter, in order to aid further extrapolation of the effects of increasing corpus size.

2.5. Using a Large Corpus

Although there are several large collections of electronic text available, investigating their exploitation is somewhat problematic, especially if one wishes to elicit information on how the increase in their size has affected the researchers who examine them. The problems stem from two main sources: firstly, most of the corpora are static, making it difficult to carry out a thorough-going investigation of the effects of their increase in size and secondly, there is no easily identifiable user base for these resources – the BNC for example is readily available in the UK, yet few researchers have succeeded in exploiting it because of the technical difficulties involved in setting it up. What is required, then, is a group of people who regularly use corpus data in a professional capacity and where the amount of corpus material has significantly increased in size. One such group of users is to be found at Cobuild, HarperCollins' Birmingham-based EFL reference book production unit. Cobuild belongs to one of the largest publishing empires in the world, since HarperCollins is in turn part of Rupert Murdoch's collection of companies. This gives the corpus builders at Cobuild access to vast amounts of text of many different varieties and has enabled them to build up a corpus of around two hundred million words, which HarperCollins has named the 'Bank of English'. Eight of the Cobuild staff kindly agreed to assist this study and the following sections provide an insight into the tools and

working methods of the lexicographers and grammarians, based on personal communications over several months in early 1995, when the Bank of English stood at around 170 million words.

2.5.1. About the Respondents

Since Cobuild is in the business of producing up-to-date dictionaries, the corpus continues to grow, as more recent material is added to it. The corpus users at Cobuild are in effect witnessing on a regular basis the phenomenon of corpus growth † and as such are in an ideal situation to describe the advantages and disadvantages of manipulating such large amounts of corpus data. Many of them worked using the corpus when it had attained only a fraction of its current size, around twenty million words, and so, it was hoped, would have an interesting perspective on the comparative usefulness of the corpus as it has grown in size over the past few years.

All the respondents used the corpus every day and three quarters of them were lexicographers, engaged in dictionary or dictionary-related projects; the remaining two were grammarians, working on the Cobuild 'Verbs' Pattern Grammar. For the lexicographers, the motivation for using the corpus differed depending on whether the task in hand was compiling a dictionary entry from scratch or revising an existing one or editing an entry which had just been compiled by another lexicographer. All the corpus users understandably made reference to the need to reflect 'what really happens in language' (i.e. instead of relying on intuition) and to the need to provide suitable examples for inclusion in dictionaries.

When asked about the issue of examining large numbers of concordance lines, the majority of respondents stated numbers in the hundreds as being a manageable amount, with two going up to 1,000 and just one considering 3,000-4,000 still feasible. There was a general trend not to look at all the available data, but rather to take a randomly-selected

† In late 1997, 90% of Cobuild staff were made redundant and the operation scaled down to a minimum. Further development of BoE appears to have been put on hold.

subset of the concordance lines. The response to this question seems to depend to a large degree on the working habits of the individual corpus user. One respondent made the point that if one is just looking at the node word, then it is preferable to look at more lines (in order to get an overview), whereas if one has already narrowed down the concordance by searching for a node plus collocate, then fewer lines are required (since one is probably seeking confirmation of the existence of a feature, rather than searching for new linguistic phenomena).

For the Pattern Grammar team, the main task was to verify whether a pattern exists at all, rather than to see how frequent it is. This therefore involved searching for specific verb plus complement combinations. This 'existential' approach was particular to this project, however, since the dictionary editors' aim is to identify the patterns, not to check for the existence of predetermined ones.

2.5.2. Overview of Methods

The Cobuild team has been using corpus data for many years now, which has resulted in the evolution of some carefully refined software and techniques for extracting information from the corpus. All the respondents use a combination of these techniques in order to achieve their aims, so let us first describe the individual techniques.

Sorting

This technique was discussed in the previous chapter as one of the direct uses of concordance lines. It involves taking a set of concordance lines and resorting them on a particular column of words, relative to the node word. This might be 'one to the right' or 'two to the left', for example. The effect of this is to place all occurrences of particular word types together.

Exactly which column is selected depends to a considerable extent on the type of word being examined. For a verb, one might want to sort on the word to the right of the node, in order to determine which are the most common complements, whereas for a

noun, sorting to the left might be preferable in order to identify the most frequent pre-modifiers.

This is frequently the first type of corpus analysis that is applied, since it can quickly offer a overview of the more obvious features of the node word, more so because the efficiency of the corpus software means that the sort column can be rapidly changed, allowing the experienced corpus user to briefly scan the different columns, looking for interesting phenomena.

Random

In a large corpus, where it is quite likely that the word under investigation occurs many hundreds of times, one possible approach is to look at only a subset of the occurrences, randomly selected from the total. The advantage of this method is that frequently-occurring phenomena will be more easily identifiable, since the less significant phenomena are unlikely to be present in the random sample. This has the effect of making the frequent patterns stand out more prominently, especially when this technique is combined with sorting. The disadvantage of this approach is that it is quite possible to lose a particular phenomenon altogether, since the selection of lines is entirely random. Supposing that a collocate occurs with its node 10 times and that the node occur 5000 times – if the corpus user takes a ten per cent sample, 500 lines, then, according to probability, there will only be one occurrence of the collocate in the randomly selected lines. Since the selection is random, there may be more than one, but there may also be none at all. It is thus possible that a reasonably significant pattern would be missed entirely using this method.

Collocate List

Cobuild's corpus retrieval software offers the facility of generating a list of the most significant collocates of the node word under scrutiny. The methodology for this kind of analysis was described in the chapter on the use of concordances in section 1.3.1. The result is a list of words which regularly occur in the context of the node word more

Collier

frequently than would be expected by chance alone. This list acts as a useful pointer towards the collocational patterns in which the node is involved, but as stated in the description of this facility in the previous chapter, it has several drawbacks which make it nothing more than an approximation to the collocational behaviour of the node.

Picture

This approach, developed in part by the author, was introduced in Chapter 1 under the heading 'Columns'. It has been included as a facility in the Cobuild software suite, where it has acquired the name 'picture'. This means of analysis attempts to address the shortcomings related to the random sampling and collocate list methods. It is an automatic methodology for identifying the most significant collocates of the node word in question, where each column is differentiated, resulting in a list of the significant collocates of the node *for each position in its context* – referred to as the node's *picture*. If one were considering a span of four words either side of the node, then the picture software would produce eight lists of collocates, one list per slot to the left and right. These lists can be ordered on various statistical measures of significance or association.

To recap, the advantage of 'picture' is that it is far more comprehensive than random sampling, making it far harder for the corpus user to miss significant collocates. The major drawback of it, however, is that it makes no attempt to look for patterns *across* the columns of collocates. This means that a collocational phenomenon can only be identified by means of one of its constituent collocates.

Regexp

Regexp or 'regular expression' is a means of specifying a search pattern in order to reduce the number of concordance lines to be analysed. It is usually used as as follow-up method after having looked at either a randomly-selected subset of lines or output from picture. The implication is that the corpus user already has an idea that a particular phenomenon exists and wants to confirm this. For example, picture output might

Collier

indicate that 'drop' is a significant collocate of 'zone'. It would then be possible to look at only those lines where 'drop' and 'zone' occur together. It is equally possible that a particular node-plus-collocate pattern has been identified in a random sample and that the corpus user then wants to see all the occurrences of it.

Regexp search may also be used inversely in order to *exclude* lines which match a particular pattern. In this way it is possible to perform an analysis of part of a set of concordance lines and then remove the analysed lines in order to concentrate on the remainder. An example of this might be that one calls up the lines for 'zone', analyses those lines containing 'drop zone' and then eliminates them by means of an inverted regexp.

This method can also help to overcome picture's lack of ability to recognise patterns which span several columns. If picture is run on a set of concordance lines which were selected on the basis of the node plus a collocate, an implicit cross-column feature (the node and its collocate) is identified, which can then possibly be expanded upon by the outcome of picture. To return to the 'zone' example, let us suppose that the corpus analyst has identified the pattern 'exclusion zone' (possibly even by means of picture). By limiting the lines passed to picture through a regexp search for 'exclusion zone', it should become apparent that there is a significant collocation with 'total', thus revealing the larger feature 'total exclusion zone'.

Word + Word

This method reduces the number of lines presented to the corpus user by stipulating that at least *two* target words should be present in each concordance line. This is generally used to isolate a node and one of its collocates, so one might search for 'war + zone', instead of simply 'zone'. This method usually makes use of the corpus indexing system to automatically find locations where the two words co-occur, and is therefore generally slightly faster in response than the regexp method, which searches already-

existing concordance lines for occurrences of the search term. The two search words are not necessarily adjacent: one might also search for "'blow' within four words of 'up'", in order to find examples of a phrasal verb such as 'blow up'.

2.5.3. Strengths and Weaknesses of the Corpus Access Methods

2.5.3.1. Random Sampling

The issue of greatest concern among the Cobuild corpus users was that the application of analytical methods which in any way reduce or summarise phenomena present in the corpus raises the possibility of some features being overlooked. This might happen, for example, if they fall outside a random sample or below an established statistical threshold. The facility which is most prone to this shortcoming would appear to be the random sampling method, for the reason described earlier: a concordance line which is not included in the sample might as well not exist.

The advantage of random sampling is that it can save much time, since it can reduce an unmanageably large number of concordance lines to an arbitrary size to suit the preferences of the user. These preferences might be expressed in terms of the amount of time available to perform the analysis, or the maximum number of lines which the corpus user can manipulate directly.

It can happen that the concordance lines which are presented as the result of a random selection do contain some feature which the corpus user wishes to exemplify, yet none of the lines are suitable, for one or more of the reasons which are discussed in Section 2.6. It is then necessary to perform an explicit search for the feature in a larger sample or in the corpus as a whole in order to find an example which is suitable for inclusion. Naturally, this takes time and detracts from the time-saving aspect of the sampling approach.

2.5.3.2. Picture & Collocates

When using those tools which transform the concordance lines into some kind of summary, such as 'picture' or 'collocates', a threshold of statistical significance is applied which tends to disallow features which have a low frequency. While this is very useful for gaining an overview of the key features of a node word, it can be a disadvantage if one is looking for new phenomena, which tend to occur infrequently when they first appear in the corpus. Since one of Cobuild's aims is to be as up-to-date as possible, this drawback is of some concern to their corpus researchers.

An issue related to this has to do with the fact that these tools summarise what is happening at the *lexical* level only. Because of this, 'picture' may fail to detect paradigms, where many words may fill a particular slot in the node word's context, since each word may not occur frequently enough to be shown as significant.

Analytical methods which present an abstract or summarised version of corpus data can also be information-losing in a way which clouds the corpus users' perspective on a particular feature. This can happen when a phenomenon occurs mainly in a sub-corpus, such as 'spoken material' or 'newspaper data'; it is difficult to convey this directly in 'picture' or 'collocates' – the corpus user has to actually call up the lines containing the feature, at which point the source is revealed.

2.5.4. Characteristics of a concordance line affecting its inclusion as a dictionary example

2.5.4.1. Positive Features

When the respondents were asked why they selected particular concordance lines in preference to others, they put most emphasis on the concrete attributes of the concordance line, attributes which, to a large extent, could be automatically isolated. The majority of respondents, lexicographers and grammarians alike, identified the presence of strong

collocational patterns as the major key in selecting concordance lines. Several of the respondents favoured lines which were brief, two of them adding that they preferred ones which were conceptually or syntactically self-contained. Clarity was an attribute which was valued by three respondents, this being manifested in a preference for those lines which clearly showed a particular syntax or collocate pattern and did not need to be edited before being included in the reference text. The need to exemplify a syntactic pattern was raised explicitly by three respondents.

2.5.4.2. Negative Features

The negative features tended, interestingly, to be concerned with more meta-textual and extra-textual features, which contrasts markedly with the more physical characteristics identified as being positive in nature. The feature most likely to cause a line to be rejected is if it is offensive in some way, in that it contains either obscenities or offensive or contentious references to real people. Indeed, references of any nature to people or places will generally be ruled out, not only because of the possibility of causing offence, but also because these tend to set the example in a particular cultural or chronological framework. Since Cobuild products sell in many countries of the world, cultural references are generally avoided, and proper names can easily cause the text to date rapidly, if, for example, a prime minister's name is used in a definition and then there is a change of government.

The above features are largely extra-textual, depending upon the sensitivity and real-world knowledge of the corpus user to identify them. Several negative features which depend on the concordance lines and the corpus which contains them were also mentioned and included the use of obscure vocabulary, where obscure means low in frequency in the corpus as a whole, or limited in some way to particular genres or varieties of English. This was made more explicit by some respondents, who said that they rejected lines which contained nonce formations. However expressed, though, this is obviously indicating that an evaluation of the typicality of the context represented in the

concordance line is being carried out by the corpus researchers.

2.5.5. Issues Relating to Increased Corpus Size

Some corpus users felt that the increase in corpus size would continue to be beneficial, citing the advantages it would bring in being able to find more and better examples of low frequency items and to separate out specialist meanings from slightly more frequent words where perhaps there is currently insufficient evidence to create a dictionary entry for a new sense. One respondent went as far as to theorize that a corpus of three or four times the current size would be capable of yielding up any desired example, presumably meaning that a corpus of around a thousand million tokens would approximate to the 'universal' corpus discussed in Chapter 1. Along with the sense of the benefits of increasing corpus size however, came an awareness among the users that the expanded corpus would require even more time to analyse.

Several respondents expressed concern regarding the configuration of the corpus, stating that the corpus as a whole would be unwieldy (since it already is to some extent) and that some kind of useable sub-division was needed. The content of the corpus would also be important. The balance of text-types might be affected if the corpus were to be enlarged using the more readily available types such as newspaper data, at the cost of the more labour-intensive data like spoken material.

Another major implication of the increase in size of corpora would be the impact on corpus access. Specifically, the speed of access, the time taken by the software to retrieve and display the information requested by the user, may be degraded, especially for complex searches on frequent items. In addition to the response time issue, reservations were expressed regarding the ability of existing methods to deal with the large amounts of output that would be generated by a much-expanded corpus. It was felt that there would be a need for new methods of looking at large numbers of lines and new sampling methods which worked more intelligently than the current system. As one respondent pointed out,

with growing corpus size, a fixed-length sample of, say, 500 lines, becomes increasingly less representative as the ratio between the size of the sample and the total frequency of the word being sampled decreases.

2.6. Conclusions

2.6.1. The Rôle of Collocates

In the responses which addressed analytical methods and positive features of concordance lines, much evidence was presented supporting the importance of the collocate profile of the node word, firstly in determining its typical behaviour and secondly in assessing whether a specific line adhered to that behaviour. This is of direct relevance to the work in hand, since the analytical approach employed by the software described herein relies entirely on information extracted from the context of the node word. In as much as they provide a summary of the contexts of the node word, collocates are an accessible, if somewhat simple, abstraction. To a large extent this is due to the limitations of the computational and statistical approaches involved in collocate identification. The fact that the KWIC format has persisted from printed to VDU form and the extensive manual analysis that has been carried out on it suggest that the concordance is not just a computationally convenient format, but rather that it sufficiently encompasses the context of its node word in its own right.

We have, however, seen that the current automated methods of establishing the collocate profile of a node word entail certain compromises, in particular with regard to the identification of collocational patterns which cross several columns or which can occur in a variety of positions. The ability of any analytical software to identify and record the fact that a feature occurs in specific positions would therefore seem to be a crucial facet of its design if it is to perform automatically a successful analysis of a node word's collocational behaviour.

2.6.2. Representativeness of the Corpus

Since one of the most important reasons for using corpus data is to provide an accurate description of the language, the corpus itself should be as representative as possible, so that the value of any analysis based on it, either manual or automatic, is maximised.

2.6.3. The Need for Filtering

Several respondents stated explicitly that there is already a definite need for more sophisticated filtering software. As the size of corpora continues to increase, this need will be even greater. The effect of the continued growth of corpora will be illustrated in the next section with concrete numerical examples of the problems which face the users of the large-scale corpus, set against a historical background of earlier, smaller corpora.

2.7. Problems with Large Corpora

The huge amounts of corpus data which are now available will inevitably lead to problems for those charged with turning it into useful information. Corpus researchers, who generally work online, can look at a screenful of roughly thirty concordance lines at a time. From that they need to gain a sufficient overview, if not of the entire node word, then at least of some subset of its occurrences. Let us suppose that for a given corpus, a screenful (or a few screenfuls) delivers a sufficiently representative impression of the behaviour of the node word under scrutiny. The growth of any corpus has the effect of increasing the frequency of many of the word types that it contains. When examining a corpus consisting of many million of tokens, therefore, the number of lines to be analysed can grow from hundreds to thousands. When this happens the corpus user's window on the data becomes smaller, since the window is never enlarged, but the amount of data to be considered grows constantly, meaning that the user needs to look at more screenfuls and the difficulty of gaining an overview becomes greater. One consequence of this development is that corpus software must be made more powerful and complex with each

generation of corpora. This is noted by McEnery and Wilson, who highlight 'the pressure put on retrieval software forced to deal with presenting thousands of answers' (McEnery & Wilson 1996 p 174) and go on to underline the need for 'smart' software, as well as mentioning the system described in this thesis as a further aid in using large corpora. The problems associated with the ever-increasing size of textual databases have not gone unnoticed by those researching the techniques of using corpora; Biber et al remark:

> Concordances are an important aid to lexicographers in identifying the various senses of a given word, and they represent a major advance over the manual sorting of citation index cards (still practiced in some lexicographic organizations). Since manual techniques depend on skill and coverage of human readers, there is no assurance that all major senses of a word will be represented; further, manual techniques provide no reliable basis for assessing the relative frequency of different word uses. In contrast, concordances based on large corpora can provide too much information, so that lexicographers are overwhelmed by the amount of data. For example, the concordance for *certain* extracted from a 10-million word sample of the Longman/ Lancaster Corpus contains approximately 3,000 entries. Simply identifying the major patterns in a database of this size is a daunting task; to group different uses accurately and rank them in order of importance is not really feasible without the use of additional tools. (Biber et al 1994 p 172)

As we saw in Chapter 1, the working methods of the researchers have evolved to take advantage of the technological benefits now available, such as on-the-fly concordance generation and instant re-sorting of the concordance lines. It is now expected that these operations take place online at the computer screen, even though this may display far fewer lines at a time than the original hardcopy concordances. As we move into the age when corpora are measured in thousands of millions, the feasibility of looking at the data through such a small window will be diminished even further. In Clear (1995) we find this very problem expressed by one of the builders of the Bank of English, the corpus referred to here:

> One obvious side-effect of using such a large corpus is that if lexicographers are to describe comprehensively the English which is evidenced there they will require ever more sophisticated software to assist with the sorting, sifting and evaluation of the mass of data. (ibid p. 1)

The author quoted here was also, incidentally, involved with the design and creation of the BNC, and therefore has much experience in large-scale corpus building projects.

In order to illustrate more concretely the problem posed by ever-larger corpora, let us look at a specific example. In the previous section, the results of a survey of regular users of corpus data were presented. One of the goals of that survey was to ascertain the maximum number of concordance lines which could be examined without the corpus users losing their overview. The responses ranged from several hundred lines up to around 3,000, or between ten and a hundred screenfuls. If we take as our maximum a figure of 1,000 lines, based on the responses from the Cobuild corpus users, and apply it to the frequency list of some well-known corpora (see Glossary for details), we may begin to get an impression of one of the disadvantages of very large corpora.

The table below lists, for each corpus, the number of types which occur more than 1,000 times.

Corpus	Total tokens (M)	Total types	N types F>1000	% types	Equiv tokens	% total tokens
LOB	1	57,420	106	0.18	570,350	53
BCET	20	217,508	1,784	0.82	14,311,798	71
BNC	99	679,525†	7,623	1.12	86,192,376	87
100M	105	628,151	7,812	1.24	94,084,063	89
BoE	211‡	638,901	12,432	1.90	196,704,042	93
ACR	435	1,127,021	19,210	1.70	413,090,408	95

Table 2.1
Number of Types with Frequency over 1,000

The first thing to notice here is that as the corpus gets larger, the percentage of frequent types increases. This is not unsurprising, since as the number of tokens grows, one would expect to find more occurrences of already-known types, many of which will pass through the 1,000 threshold. This is emphasised by the fact that the introduction of new types slows down as the corpora get bigger; thus although BCET is twenty times the size LOB, it only contains about four times as many types; BoE likewise is ten times larger

† Upper and lower case forms are counted separately, hence the slightly higher type count for this corpus.

‡ Note that the BoE had by now increased in size from 170 million to 211 million.

than BCET, yet its type count is only three times greater. ACR differs slightly from this pattern, having twice as many tokens as BoE and also nearly twice as many types. It must be borne in mind, however, that ACR is composed entirely of newspaper text and is therefore likely to contain a very large number of types, since the language in newspapers is arguably the most rapidly evolving, generally being the first printed medium to present references to new real-world entities.

More striking are the columns relating to the total tokens which the frequent types represent. This is calculated by finding all those types which occur 1,000 times or more and summing their frequencies. In the BCET, 1,784 types have a frequency greater than 1,000, which corresponds to only 0.82 per cent of all the types (217,508). In terms of tokens, however, this corresponds to 14,311,798 running words, or 71 per cent of the entire corpus.

The situation worsens as the corpus size expands: the massive size of the 211 million-word Bank of English (BoE) pushes 93 per cent of its content over the frequency threshold and the 435 million-word database of newspaper text (ACR) raises this to 95 per cent.

In terms of useability of the corpus, this phenomenon means that the majority of the evidence stored in the corpus is not available to corpus researchers using straightforward concordancing as a means of analysis. They must therefore find other means to get at the information embedded in the corpus. To return to Clear:

> ... it is clear that for many words in the central core of English vocabulary items the sheer frequency of occurrence of these words makes it very time-consuming to carry out an analysis of concordances without further software assistance.(Clear 1995 p. 10)

Several different methods for the automatic analysis of output from a corpus were introduced in the Chapter 1. It was stated then that they each have particular shortcomings, although the majority of them served a useful purpose in reducing the amount of information presented to the corpus user, so that some kind of overview could be obtained of the behaviour of a given node word. The way in which the reduction is carried out, however,

is the key to the problem with these analytical methods, since it is is not performed intelligently, being based instead on random numbers, or the occurrence of a particular collocate. This chapter has shown, based on evidence from full-time professional users of a large corpus, that the potential shortcomings of the various analytical methods are a reality.

What is required, then, is a corpus tool which uses a valid method of determining the characteristic features of 'good' concordance lines, applies that method to a set of lines in order to select the most useful ones and presents the selected lines to the corpus user in a manner which they can easily understand. In the next two chapters, some of the features of concordance lines will be examined and a comparison will be drawn between the concordance 'text' and natural-language text with a view to establishing whether an automatic analytical method developed for use on text is applicable to concordances.

Chapter 3

Concordances and Abridgement

3. Concordances and Abridgement

3.1. Introduction

The previous chapter outlined some of the major difficulties facing corpus researchers as they attempt to analyse the substantial amounts of concordance data which are retrievable from today's large corpora. In order to overcome the problem of information overload, it would be desirable to process a word's concordance set automatically, identifying the lines which are most indicative of the node word's behaviour and then presenting only those lines to the corpus user. One analogy to this exists in software systems which are capable of creating an abridgement of a text by extracting from it sentences which are in some way 'core' or 'key', thereby reducing the amount of information presented to the user of the system. This chapter will focus on the techniques developed for the abridgement of mainstream text, while the next chapter will examine the parallels between conventional texts and concordances and consider the possibility of extending the abridgement analogy to the point where one might literally abridge a concordance. This might lead us to think that the concordance is in fact a kind of text and that by abridging it we remove the extraneous, peripheral lines, leaving only the core, representative lines and thereby reduce the amount of redundant material confronting the corpus user.

One methodology for creating automatic abridgements of conventional texts was developed by Michael Hoey, as described in Hoey (1991) and is outlined in Section 3.3 below. His research, based on the analysis of lexical cohesion within the text, has been exploited to produce systems, both manual and computer-based, for the automatic abridgement of natural-language texts. There are a number of reasons why Hoey's system has been used in preference to any other. Foremost is the fact that the analytical techniques are in the public domain and can be and have been replicated, as in (Benbrahim and Ahmad 1995), by reference to Hoey's published research. This contrasts markedly with other abridgement systems, which, largely for commercial reasons, keep the algorithms used in their

processing hidden from the user. Examples of this are BT's ProSum service on WWW (http://www.labs.bt.com/pressoffice/archive/1997/prosum.html) and the AutoSummarize tool found in Microsoft Word. Cohesion analysis of various kinds has also been used in the creation of textual abridgements independently of Hoey's work, as in the studies by Reimer et al (1990), who examine the repetition of particular keywords and also in the work of Morris and Hirst (1991), who establish 'lexical chains' in the text which are related by cohesion.

Further advantages of Hoey's system are that it has been shown to be very flexible in its application to text of different types and languages and that the technique does not involve any complicated statistical measurements, but rather relies on counting of features in the text. This contrasts with the approach used by Salton et al (1994), who employ extra-textual word frequency data to drive statistical measures aimed at identifying the core segments of the text. In addition, Hoey's system uses sentences as input and produces a subset of those sentences as output, whereas other 'abridgement' systems in fact produce output in the form of keywords (Källgren 1988) or some higher-level abstraction of the text structure (Morris & Hirst 1991).

The identification of cohesive features is not simple, but the simplicity of the overall technique developed by Hoey makes it more readily implementable and verifiable on a computer, even if the range of automatically identifiable features, as we shall later, is somewhat reduced. No reference is needed to external sources of information or to characteristics of the text other than the degree of cohesion, such as sentence length or type-token ratio. In addition, no thesaural expansion is carried out on the text, in contrast to the methods used by Benbrahim & Ahmad (1995) and Källgren (op cit), which further assists in the verification of the results.

The remainder of this chapter will examine the ways in which lexical cohesion has been used to produce automatic abridgements of normal text and in the subsequent chapter we will investigate the feasibility of applying this methodology to the concordance by means

of a comparison of some of the key features common to concordance lines and sentences within a text.

3.2. Text and Lexical Cohesion

Lexical cohesion was identified by Halliday and Hasan (1976) as one of five classes of *cohesive ties*, the others being *conjunction, reference, substitution* and *ellipsis.* These 'ties' represent those relationships between features of the text which help to organise it and make it coherent and cohesive. (See Hoey, 1991, p.11 ff for a discussion of the important distinction between coherence and cohesion). It may be of interest to note here some of the observations which Halliday and Hasan made concerning the nature of 'text'. They define it as

> ... any passage, spoken or written, of whatever length, that does form a unified whole. (Halliday & Hasan, 1976, p. 1)

Furthermore, a text is said to be

> a unit not of form but of meaning (ibid p. 2)

The point to note here is that if we follow the analogy presented earlier then the 'text', that is, a concordance, does not correspond to either of these definitions.

A concordance forms no 'unified whole', since its content depends firstly on the presence of a particular lexical item (the node word) and secondly, and more indirectly, on the constitution of the corpus from which it is drawn. This second factor enables the 'text' of the concordance potentially to include *both* spoken and written material simultaneously, given that both types are present in the corpus as a whole.

As to the second definition, it is clear that a concordance is not a unit of meaning – it cannot, for example, be read as a whole – but rather that it is circumscribed by the number of characters the corpus user has chosen to include in each line. It is in fact defined in terms of its *form*, having, in the case of the KWIC format, the node word in the middle of each line, with a fixed number of characters of context either side of the node.

Halliday and Hasan go on to define what is meant by *cohesion*. They say that it

> refers to relations of meaning that exist within the text, and define it as a text. (ibid p. 4)

As we saw earlier, there is no semantic relation between the constituent elements (the lines) of a concordance, only a formal one. Is it, then, possible to refer to the existence of cohesion in a concordance? Using the definitions we have seen so far it is probably not. One conclusion to be drawn from this is that we are dealing with a new type of text, one which stretches known definitions. As far as lexical cohesion is concerned, the current work does rather more than stretch its definition. In later chapters it will be shown that the ties present between the elements of the concordance have much in common with the cohesive relations between the items of a 'text' as it has been defined above.

3.3. Abridgement

Halliday and Hasan sub-divided lexical cohesion into *reiteration* and *collocation* and it is this former sub-class which Hoey adopted for use in his work on textual analysis. This work eventually led him to develop the system for creating abridged versions of a text which is introduced in Hoey (1991).

Hoey noticed that the sentences of a text which had a large degree of cohesion with other sentences could be used to create an abridgement of the text. He developed a methodology which enabled him to record and utilise the presence of cohesive ties, which he termed *links*, to apply a measure to each sentence as to how much it contributed to the cohesiveness of the text. This measure was expressed in terms of the number of *bonds* the sentence obtained. Hoey defined a bond as 'a connection which exists between a pair of sentences by virtue of there being an above-average number of links relating them'.

At the simplest level, the cohesive relationship between any pair of sentences was expressed in terms of the number of links between them, with several different kinds of link being possible. In summary, these were:

Simple (Lexical) Repetition

The link exists between two instances of the same word, for example 'dream' ↔ 'dream'.

Complex Lexical Repetition

Here a link is made between members of a lemma, e.g. 'dream' ↔ 'dreaming'. This category also includes morphologically related antonyms of the 'happy ↔ unhappy' variety.

Simple Paraphrase

This instantiates a link between synonyms which are coreferential within the text ('clever' ↔ 'intelligent').

Complex Paraphrase

This covers some cases of repetition by antonym not covered by complex lexical repetition, such as 'light ↔ dark'.

Hyponymy

Here, superordinates are included as repetitions, e.g. 'tulip ↔ flower'.

Pronominal Repetition

In this case, reference is made to a lexical item by means of a pronoun and a link is therefore made between the pronoun and the lexical item, e.g. 'dream' ↔ 'it'.

Substitution

Where a phrase such as 'the first one' or 'the above' is used to refer back to a previous clause, sentence or even larger section of the text, this must also be considered as a link.

Hoey's approach involved the creation of a two-dimensional *matrix*, of which only the lower half, below the diagonal from top-left to bottom-right, was used. The cells of the matrix were referenced by two numbers, each referring to the numbers of the sentences in the text. Thus cell (2,5) would contain information pertaining to the relationship between

sentence 2 of the text and sentence 5. Hoey analysed the text by identifying instances of each of the phenomena listed above. Whenever he discovered a link between two sentences, he incremented the appropriate cell of the matrix. A sample matrix, constructed by this means and based on Hoey's own example, is given below. The figures in parentheses identify sentence numbers; cell values in square brackets indicate questionable links.

```
     (1)
(2)  6    (2)
(3)  2    1    (3)
(4)  5    1    2    (4)
(5)  1    0    1    0    (5)
(6)  3    1    2    1    1    (6)
(7)  4    0    3    2    2    2    (7)
(8)  5    1    2    4    0    1    1    (8)
(9)  1    1    0    2    0    0    0    0    (9)
(10) 2    0    0    2    0    0    1    2    1    (10)
(11) 1[3] 0    3    1    1    2    2    1    0    0    (11)
(12) 3[4] 0    1    3    0    1    2    1[2] 0    2    1    (12)
(13) 0    0    0    0    0    0    0    0    0    0    0    1    (13)
(14) 2[3] 0    1[2] 2[3] 0    [1]  2[3] [1]  0    1    [1]  3    0    (14)
(15) 0    0    0    0    0    0    0    0    0    0    0    0    0    2    (15)
(16) 1    0    0    1    0    0    1    1    0    1    1    2    0    3    3
```

Figure 3.1: Sample Matrix (from Hoey 1991 p 90)

Once the entire text had been processed in this fashion, a link threshold was applied to the matrix. This was generally set to the value **3** and was applied to the connectivity data recorded in each cell of the matrix in turn. Any cell which contained a value greater than or equal to the link threshold was deemed to have established a bond. That is, if two sentences *S1* and *S2* had three or more links, then those sentences were said to be bonded. The value of three as the link threshold was determined heuristically by Hoey on the basis that it is high enough to prevent 'accidental' bonds from being formed where sentences happen to have words in common, yet low enough to allow the majority of valid bonds to occur.

Having applied a link threshold in order to translate links into bonds, the next step was to use the bonds to decide which sentences were to be included in the abridgement. This was achieved by reference to the number of bonds acquired by the sentences of the text.

Because of the triangular shape of the matrix, it was necessary to read both across the row and down the column for a given sentence number; by identifying the number of sentences with which a bond had been formed, based on the link threshold discussed previously, a bond score was established for each sentence. In order to determine which sentences to include in the abridgement a further threshold, the *bond* threshold, was implemented, such that any sentence whose bond score attained the threshold would be included in the abridgement. Supposing a bond threshold of **3**, any sentence which shared a bond with three or more other sentences would be selected to form part of the abridgement. In the sample matrix above, sentence (12) shows an example of this: reading across the matrix row it forms two bonds, one with sentence (1), the other with sentence (4), both of which have at least 3 links to sentence (12). Reading down the matrix column we find one further bond, formed with sentence (14), since it also has three links to sentence (12).

In this chapter we have examined one means of creating abridgements from conventional texts by selecting key sentences from them according to the number of bonds that they acquire when analysed for lexical repetition. If this same methodology is to be applied to the concordance 'text', then it would be appropriate to identify any features of conventional text and concordances which may be common to the two text types. In the next chapter, therefore, the suitability of concordances as candidates for abridgement will be explored.

Chapter 4

Concordance Lines as Text

4. Concordance Lines as Text

4.1. Introduction

In the previous chapter we introduced the hypothesis that the concordance could be treated as a new kind of text and that it would be reasonable to apply a text-based abridgement technique, Hoey's lexical cohesion analysis, to the concordance as though it were a conventional text, the aim being to select from the concordance those lines which contained features which are in some way key or central to the characteristics of the node word under scrutiny. What is proposed, then, is to apply the abridgement methodology to sets of concordance lines, such that the lines are treated as the 'sentences' within the concordance set 'text'. As a means of testing the hypothesis that concordances have text-like features and can thus be 'abridged' it seems appropriate to look at some of the features which are common to sentences and concordance lines.

In the sections which follow, the umbrella term *elements* will be used to refer to both sentences and concordance lines. When we are dealing with a natural-language text, an element corresponds to an orthographic sentence. When concordances are the target of the analysis, an element consists of a single concordance line. There are a number of parallels between these different types of element, which will now be discussed.

4.2. Size

There is a strong relationship between an element's size and the number of lexico-cohesive links which it can potentially form with other elements.

4.2.1. Sentences

The size of a sentence, defined in terms of the number of tokens (running words) it contains, varies considerably. A preliminary analysis of a sample of the Financial Times from January to March 1994 inclusive indicated that there is a great variation in the possible

length of sentences. Sentences containing between one and forty tokens accounted for about 95 per cent of all the sentences, with the most common length being six words, although several sentences were found to contain over one hundred tokens. The sample was 6.91 million tokens in size and contained approximately 364,000 sentences, giving an average sentence length of around 19 words. More recently, the analysis has been repeated using the written component of the British National Corpus (BNC), which includes sentence delimiters as part of its SGML markup. The written texts account for approximately 90 million out of the total 100 million tokens in the corpus and are divided into 5,188,373 sentences, or 'S-units' to use the BNC terminology, giving a mean sentence length of 17.3 words or 'W-units' (Burnard, p 7). My own analysis, based upon the SGML tagging in the written texts, indicates that the most commonly-occurring length of an S-unit is five tokens, although the difference in frequency among the commonest sentence lengths is minimal. While the scope of an S-unit in the BNC includes text which occurs in headlines and titles, an examination of these does not seem to indicate that the length distribution is significantly different to that found in the texts themselves. The figures for the top ten sentence lengths can be seen in the next table.

Rank	Sentence Length	Number of Occurrences
1	5	193,565
2	3	193,377
3	2	192,807
4	4	191,583
5	6	191,554
6	7	188,435
7	8	181,904
8	1	176,259
9	9	174,717
10	10	170,819

Table 4.1

Commonest Sentence Lengths in the BNC

At just over 10 million words in total, these sentences account for 1,855,020 out of the 5.2 million sentences, or about 36%. In terms of tokens, they represent approximately 11% of the corpus. While the most common sentences are relatively short, several surprisingly long sentences were also found in the BNC, with 3,302 S-units containing 100 W-units or more, equivalent to about 0.5% of the written component. In total, 300 different sentence lengths were found.

It is interesting to note the closeness in size of the frequencies of occurrence of the most frequent sentence lengths, such that from the most frequent sentence length to the 10th-ranked length there is a fall-off of only 12% (193,565 – 170,819).

The next table shows how the frequencies for the 10% centile ranges of sentence length are distributed. This type of analysis is intended to convey an impression of the fall-off in frequency as the sentence length increases. The table was arrived at by building a list of all the sentence lengths present in the corpus and sorting the list into descending order of the frequency; in effect this would resemble the above table with all the items from rank 11 to rank 300 added. The 300 items on the list were then divided into ten sets, such that

the first set, the 90th centile, contained enough items to account for 10% of all the sentences. Referring back to the previous table, this would involve only the first three items, which have a combined frequency of 579,749, equivalent to 11.1% (rounded to 10%) of the 5,183,719 total sentences. The remainder of the table which follows was built up by adding more items until 20% of the sentences were accounted for, resulting in the 80th centile score and so on until all 300 items have been included, giving the '0th' centile. The slightly unusual feature to note here is that the first six centile ranges all add the same number of items, three. This reflects the closeness in frequency of the commonest sentence lengths which was noted earlier, indicating that the trend exemplified in the first ten items continues at least as far as the 18th item and it is not until the range below the 20th centile is entered that more than 5 items need to be added to account for the next 10% of the total number of sentences.

Centile	No. of Items	No. of Sentences
90th	3	579,749
80th	6	1,151,321
70th	9	1,684,201
60th	12	2,188,974
50th	15	2,675,579
40th	18	3,138,652
30th	22	3,686,422
20th	27	4,218,558
10th	34	4,690,411
0th	300	5,183,719

Table 4.2
Centile Analysis of BNC Sentence Lengths

The frequency distribution can also be visualised by plotting sentence length against frequency. The following graph shows the frequency of sentences containing up to 100 tokens. Note the convex nature of the plot at around 20 words and compare this with the concave shape further to the right.

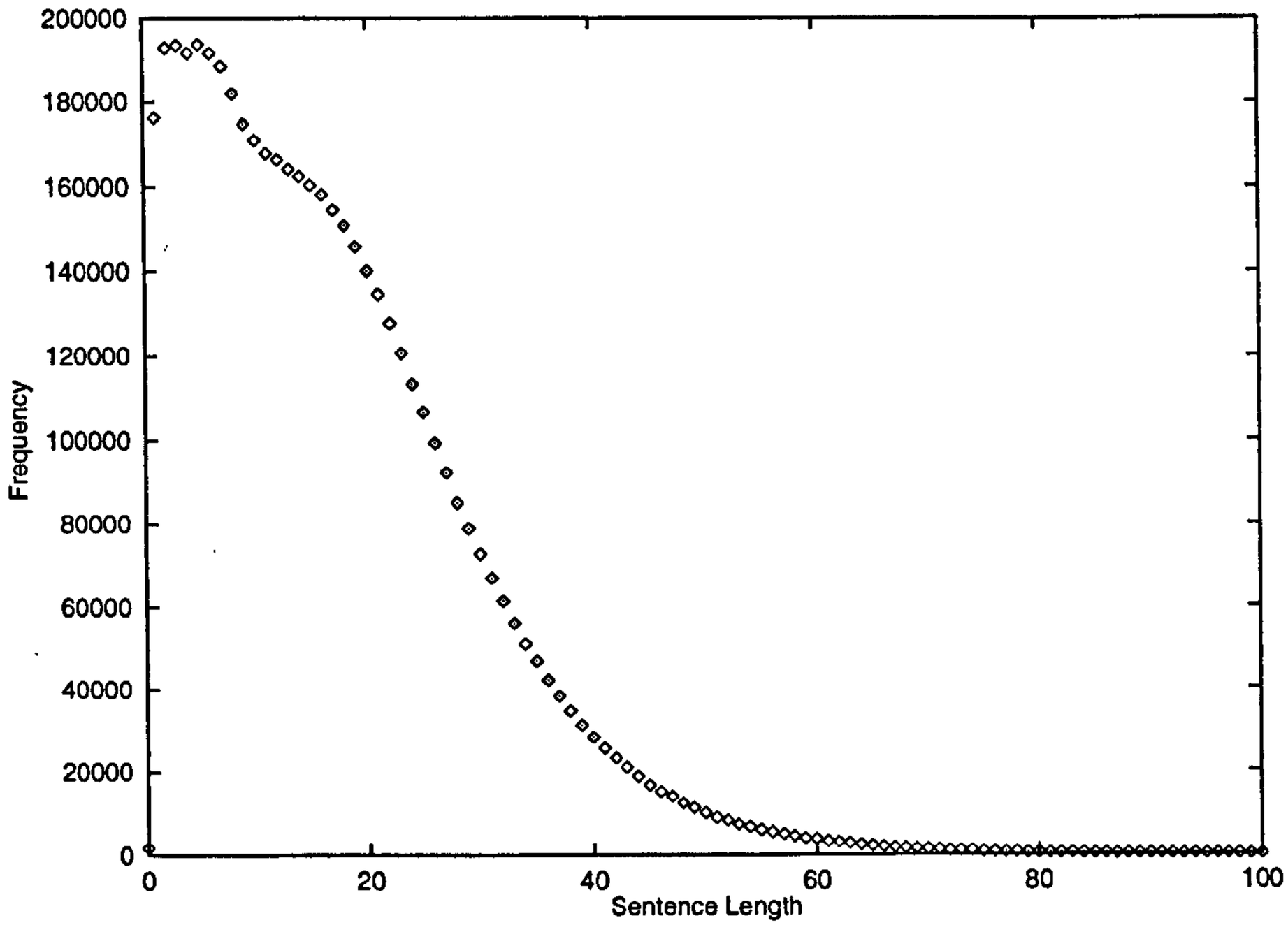

Figure 4.1
Frequency Plot for BNC Sentence Lengths

4.2.2. Concordance Lines

Concordance lines in Keyword-in-Context (KWIC) format also vary in the number of tokens that comprise them, but, due largely to their fixed width they vary far less than sentences. The concordance lines analysed for this study are 80 characters wide and contain between eight and twenty-nine running words, the most common length being fifteen words. There are several reasons for the use of a fixed-size concordance line. Firstly, *cohort* was intended to be as general-purpose as possible – most concordancing packages are capable of producing a text file containing the user-selected concordance lines, which could then be fed directly into *cohort*. Secondly, much of the data on which *cohort* was trialled was only made available as text files of the concordance and had to be transferred across several different computer systems of varying platforms in the course of this study. There was therefore no scope to make use of the extended contextual information which is provided by interactive concordancers such as WordSmith Tools or TACT. And finally, most corpus users' point of initial contact with a word is the simple, unexpanded

concordance, one line of context per occurrence, as wide as the VDU will allow. The word count characteristics of several concordances are summarised in the following table, which contains figures for the number of tokens found in each line of all the concordance sets analysed, a total of 29,233 lines.

Number of tokens in line	Raw Frequency	% of total
8	2	0.007
9	38	0.130
10	160	0.547
11	755	2.583
12	2,183	7.468
13	4,369	14.945
14	6,670	22.817
15	6,843	23.408
16	4,755	16.266
17	2,362	8.080
18	785	2.685
19	178	0.609
20	42	0.144
21	12	0.041
22	22	0.075
23	18	0.062
24	15	0.051
25	8	0.027
26	7	0.024
27	8	0.027
29	1	0.003

Table 4.3

Number of tokens in concordance lines

The 29,233 lines contain a total of 426,004 words, giving a mean length of 14.6 words, corresponding closely to the observed mode (most frequently occurring length) of 15. This would seem to indicate that the number of tokens in concordance lines lies within a smaller range than in sentences and this can be verified by calculating the variance (the mean squared deviation) † for each set of data. The results of this calculation are shown

† Establishing the variance for a set of data involves subtracting the mean length from each member of the set, squaring this value and adding it to the total. This total is then divided by the population (number of members) to give the variance. The result is a measure of the variation or range of values in the items and is generally represented by the formula:

$$V = \frac{\sum(X-\mu)^2}{N} - 1$$

in the table which follows, which, to aid comparison, also includes figures for the single concordance set for the node word 'black':

Data	Population	Mean Length	Variance
BNC Sentences	5,188,373	16.3	169.6
FT Sentences	364,144	19.0	13.4
All Concordances	29,233	14.6	3.0
'black' Concordance	6,096	14.5	1.6

Table 4.4

Variance for Sentence/Concordance Length

From these calculations it is clear that the variation in sentence length is considerably greater, two orders of magnitude in fact, than that for concordance line length.

The most important consequence of this scope for difference in size has to do with the likelihood of an element acquiring *links*. These are the means by which lexical cohesion is established and their existence relies upon the presence of lexical items which are repeated in various ways across sentences of a text. The maximum number of links which any element can achieve is limited by the number of words which it can contain and as we have seen, a sentence, with between one and, potentially, several hundred words has empirically a higher chance of acquiring a link than a concordance line, which is constrained by its physical size to less than thirty words. In addition to this, the node word is present in every line, thereby further reducing the potential for linking, since it is not eligible as a link.

It is hard to say whether one type of element is more likely to gain links than the other, since there is so much overlap between the ranges of the length of each element. If we take the mean concordance line length, 14.5, and compare it against all the possible lengths for a sentence in the BNC, we find that just under half (48.6%) of the sentences are shorter and slightly more than half (51.4%) are longer. The action of the node's

where V is the variance, μ the mean and N the number of members.

collocational† preferences must also be considered, but their overall effect will be to increase the likelihood of items occurring repeatedly across concordance lines and thereby creating links.

On the other hand, the definition of a link varies considerably, depending on which elements are being analysed. In Hoey's manual system and the abridgement software which was derived from it, only lexical words were available as candidates for linkage, but if we are dealing with concordance data, then the possibility of grammatical (i.e. non-lexical) items forming links has to be considered. If they are included as candidates for linkage, then this broadens the definition of links for concordances, since almost any repeated item would potentially take part in the linking process. To this must be added the fact that although sentences *can* be longer than concordance lines, a large proportion are not. 36% of the BNC sentences, for example, contain ten words or less; furthermore, of those words a good many will be excluded from the analysis because they are non-lexical.

What conclusions, then, can be drawn from these various comparisons? Firstly, we have established that concordance lines are more consistent in length. This is advantageous to the identification of links, since the low variance in length provides a level playing field for each line, relative to the others. This contrasts with textual abridgement, where long sentences have a greater likelihood of selection. Secondly, it would appear that the number of possible links in a concordance line is difficult to define in terms which are comparable to sentence length; the chances of a concordance line being longer or shorter than a sentence are roughly equal.

4.3. Nature of Elements

A sentence is a fully-formed unit of natural language, created with a specific communicative purpose in mind. A concordance line in KWIC format, on the other hand, is artificially truncated to a certain number of characters, possibly even severed partway through

† The term 'collocate' in this chapter is used to refer to all items which occur in the environment of a node word more often than would be expected by chance alone. It does not, therefore, exclude grammatical or 'colligational' items.

a token, depending on how it was generated. Expressed differently, one might describe the sentence as a *functional* element, whereas the concordance line, as we shall see in the chapter on the basics of lexical cohesion which follows, is a *formal* one; that is, the concordance line is defined by its form, while the sentence is defined by the communicative intent of its author.

The nature of the concordance line is that it contains a particular word, the node word, in its central position. Setting aside the issue of the format for a second, it might be said that concordance lines are the linked elements of a very large text (the corpus), where the links have been limited so that they exist solely on the basis of one lexical item – the node word. What must be borne in mind, however, is that the corpus does not consist of a series of concordance lines, rather the concordance line's KWIC format is imposed on the corpus at those locations where the node word occurs. The concordance line is not a naturally occurring unit, in contrast with the sentence.

In addition to the pre-defined link between each of the members of the concordance set created by its node word, we know that certain other links are likely, because of the existence of collocational patterns associated with the node. In the case of sentences in a text, this phenomenon is not present; there is no one feature, corresponding to the node word, which occurs in every sentence and thus it seems unlikely that the equivalent of collocates would be found there either. Where a node word does occur in a concordance line with one or more of its common collocates, the probability is high that this will contribute to the acquisition of links for that line, since the collocates are, by definition, present within the contexts of a number of occurrences of the node word. As an example, if we were to examine the concordance lines for 'black' and identified within them several lines which exemplify the phrase 'black and white' then each of these lines would be linked to the other lines containing 'black and white'.

4.4. Order of Elements

The sentences of a text must normally appear in a particular order; to present them in any other order would seriously detract from the readability of the text. Concordance lines, conversely, may appear in any order, as corpus users' understanding of one line does not depend on their having absorbed the information presented in the lines which come before it. This is perhaps attributable to the fact that they are never 'read' or 'understood' as text, but rather analysed or scanned by the corpus researcher for individual features. Perhaps the convention of displaying concordance lines sorted on a particular word to the right or left of the node is the closest approximation we have to the concept of a 'required' order of presentation. It is certainly true that an unsorted set of concordance lines is far harder to analyse than one which can be sorted and re-sorted on particular positions.

4.5. Conclusions

In this chapter the size, nature and ordering of sentences and concordance lines have been considered with a view to evaluating the applicability of a cohesion-driven abridgement approach to the problem of over-abundant corpus data.

It has been shown that there is a certain degree of overlap in terms of the number of words that each type of element can contain, although the number of links which may be derived from this is affected by other factors, related to the individual characteristics of the elements. Any automatic analysis of concordance data would therefore need to be able to take account of these differences and, where possible, exploit them in order to improve the quality of analysis.

In the next chapter we shall look at specific computational implementations based on cohesion analysis as applied to the abridgement of natural-language texts and the selection of concordance lines.

Chapter 5

The Software

5. The Software

5.1. Abridgement Software

In 1989, I implemented a software version of Hoey's abridgement procedure. This initial version was a prototype written using Unix™ tools and lacked the ability to identify any but the simplest form of link defined by Hoey, being only able to handle simple repetition, but it performed in seconds an analysis which would take a human being several hours to complete. The prototype was accepted by British Telecom Research Labs, who funded a slightly more sophisticated version, to be developed and tested in the Research and Development Unit for English Studies (RDUES) at the University of Birmingham. It was hoped that the software would become an integral part of Telecom's online textual database service, so that users would be able to see abridgements of the texts they have selected rather than having to read the whole document. The new version, written in the 'C' programming language, included the facility to identify complex lexical repetition and ran rather more quickly, but could only deal with texts of up to 256 sentences in length, because it implemented the matrix as a 256 x 256 integer array in the computer's memory, this being the largest array which could be created with the hardware available†. Since the abridgement software was initially designed and fine-tuned to work well with texts such as newspaper articles, the limitation on the number of sentences which could be processed was not considered a serious one, since few news articles are long enough to exceed the sentence limit. In the next version of the software, however, the limitation was eventually overcome by replacing the fixed-size array with a dynamically allocated memory structure which dispensed with the need to store every single cell of the matrix. A major drawback of the original matrix design, when implemented as an array in the computer's memory, was that a large proportion of the cells in the array were in fact zero, but recording the zero still required as much storage space as any non-zero value. In the

† The computer used at this time was a Sun 3/50 running the SunOS 3.5 (Unix) operating system. The physical memory was in the region of 16 megabytes, but virtual memory was also available. The array size restriction was actually enforced by the C compiler.

sample matrix given in Hoey (1991: 90), for example, 53 of the 120 cells are zero; since this is a manually created matrix, based upon all the link types identified by Hoey, we can expect that an automatically created matrix would have had even more zero cells. The use of the dynamic storage approach brought two advantages. Firstly it allowed all the zero values to be ignored, because only those cells which contained a value greater than zero needed to be stored and thus memory usage was decreased considerably. Secondly, the upper limit on the number of sentences which could be processed was removed, as the 256-item limit on the matrix no longer applied.

5.2. Selecting Concordance Lines

Once our collaboration with British Telecom had ended, the software was re-written completely, resulting in a faster, more adaptable program called **abr**. The increased capacity of the abridgement software meant that it could be run on much larger texts than had hitherto been considered as candidates for abridgement. It became possible to abridge entire books, amounting to several thousand sentences, in a few minutes.

In 1991 RDUES embarked on the AVIATOR Project (Renouf 1994, Collier 1993), the goal of which was to identify and store collocational 'profiles' for words in a large diachronic corpus and to record changes in those profiles over time. The profiles consist of a comprehensive record of the co-occurrence of the types in the corpus, storing span and frequency information from which statistically significant collocates can be derived.

It was at around this time that the 100-million word British National Corpus (BNC) was released and HarperCollins' 200-million Bank of English (BoE) began to be publicised. Informal discussions with several users of the BoE revealed some of the drawbacks of using the larger corpus and these were covered in detail in Chapter 2. As a result of these discussions, I began to contemplate ways in which the difficulties presented by ever-growing corpus size might be overcome and this, together with the increased capacity of the abridgement software and the experience of collocational behaviour in the AVIATOR

Project, led me to consider the possibility of applying the techniques of lexical cohesion analysis to concordance lines and ultimately to create a program, *cohort*, to put the theory to the test.

The primary objective of this kind of analysis was to assign to concordance lines a score of 'typicality' and present them in order, with the most 'typical' first. The need for this type of analysis is pointed out by Patrick Hanks:

> The words of English do not, typically, combine and recombine freely and randomly. Not only can typical grammatical structures and form classes be observed, but also typical collocates. The distinction between the possible and the typical is of the greatest importance. It is possible, given a reasonably lively imagination, to use a particular word in any number of different ways. But when we ask how the word is typically used, rather than how it might possibly be used, we can generally discover a relatively small number of distinct patterns, which may be used as a basis for explanations after being grouped together in appropriate ways. (Hanks 1987 p. 121)

One means of addressing the issue of typicality was created by John Sinclair and Jeremy Clear at the University of Birmingham in the late 1980s. They analysed concordances using a program called **typical**, which operates by extracting a set amount of context (generally four words from either side of the node word) from the concordance lines and creating a list of all the words which occur in those contexts (the raw collocates) from these. The frequency of occurrence of each word within the fixed context is then compared with its overall corpus frequency in order to generate a Z-score based on the observed to expected frequency. By applying a threshold to the resultant Z-scores, a list of collocates which are statistically significant is created. Once the list of significant collocates has been derived, it is matched against the original concordance and each line is thereby assigned a score on the basis of the presence of significant collocates within it.

Sadly, no written record seems to exist of **typical**, but on the basis of some minor developmental work which I did on the program, I am aware that the fundamental difference between cohesion analysis and the collocate analysis used by **typical** lies in the fact that the former system treats the concordance as a self-contained text, while **typical** functions by compiling a frequency list of all the types within a fixed context of the node and then

comparing this with the frequencies derived from an external corpus wordlist in order to identify significant collocates. It is therefore prone to the weaknesses relating to the determination of statistical significance mentioned in Section 1.3.1. In addition, **typical** takes no account of the positioning of collocates within the line, whereas *cohort* is able to do this, if required. If run on a concordance for the word 'kin', for example, **typical** might determine that 'kith' is a significant collocate. It would then apply a certain score of typicality to all lines which contained the collocate 'kith', even if it were used in an atypical phrase, such as 'kin and kith'.

In Chapter 4, a comparison of some of the characteristics of sentences and concordance lines was carried out, with a view to determining the validity of applying cohesion analysis to concordance lines. We saw how both texts and concordances are divisible into discrete elements. In terms of the approach to automatic processing described earlier, this division has two functions; firstly it defines the unit of *analysis* for the respective system, and secondly it defines the unit of *selection*. This means that each element in the input to one of the systems is analysed separately in its own right. Once all the units have been analysed, and the results collated and merged into the matrix, it is the same element which is used to generate the output, which is itself entirely made up of those elements. While the characteristics of the two elements do differ, it would seem that concordances are generally compatible with this type of approach. It is worth reminding ourselves that, as was noted in Chapter 3, there are some systems of abridgement which produce output that does not consist of the original sentences. They perform a more deep-level analysis of the input text and produce output in the form of keywords or possibly synthesised sentences. The benefit of the system employed by *cohort* is that its output is formed as a subset of its input, making it relatively simple to substitute concordance lines for sentences and produce 'abridgements' of concordances. This being the case, how might a set of concordance lines be analysed?

The starting point will be the extraction of a set of concordances for the word which we wish to analyse. Below is a sample of lines for the word 'kin', taken from BCET. This is presented as an example because it has two very strong patterns and is fairly low in frequency, while at the same time containing a line which does not fit either of the main patterns. The various stages in the automatic analysis are therefore easy to comprehend, although of course, the program would be more productively utilised on a much larger set of concordance lines.

```
(1)  barrier excuses. As for the "kith and kin" appeal, to quote the Reverend Arth
(2) id. "And I'm sure the cattle's next of kin have been informed but is that quit
(3) ted backing of France for her kith and kin in Algeria and for her Army protec
(4) earted commitment towards our kith and kin overseas." Identity of "race, langu
(5) f they do lecture their white kith and kin rather than the guerrillas, it is p
(6) e a narrow view of who is our kith and kin. Religion very properly tends to em
(7)  been seeing to that. The only next of kin seems to be a cousin in Droitwich.
(8) l kindness toward him, they're not his kin.... That's exactly the feeling. Old
(9) and her property passes to her next of kin under the  intestacy rules. That me
```

Figure 5.1
Selection of Concordance Lines for 'kin'

The two main patterns represented in this sample are 'kith and kin' in lines (1), (3), (4), (5) & (6) and 'next of kin', which is present in lines (2), (7) & (9). Line (8) stands alone in that it contains neither of these. Any automatic analysis of these lines should therefore be capable of isolating the two main features and filtering out any lines which do not contain them.

In order to establish that the above analysis could be carried out automatically, a version of the abridgement software was used in which the sentence recognition component had been modified so that, in effect, it treated each concordance line as a sentence. This component forms the first stage in the program's set of procedures and is known as the *tokeniser*, as it carries out the identification of all the individual words, or tokens, in the text, distinguishing them from punctuation, white space etc. It is also responsible for recognising the sentence boundaries. The output from this stage therefore normally consists of tokens, each one of which has attached to it the sentence number in which the token occurs. When processing a set of concordance lines then, each token is instead given the number of the *line* in which it occurred. The above concordance sample is therefore tokenised as:

Collier

```
barrier 1       is 2          identity 4    our 6         kindness 8
excuses 1       that 2        of 4          kith 6        toward 8
as 1            quit 2        race 4        and 6         him 8
for 1           ted 3         langu 4       kin 6         they're 8
the 1           backing 3     f 5           religion 6    not 8
kith 1          of 3          they 5        very 6        his 8
and 1           france 3      do 5          properly 6    kin 8
kin 1           for 3         lecture 5     tends 6       that's 8
appeal 1        her 3         their 5       to 6          exactly 8
to 1            kith 3        white 5       em 6          the 8
quote 1         and 3         kith 5        been 7        feeling 8
the 1           kin 3         and 5         seeing 7      old 8
reverend 1      in 3          kin 5         to 7          and 9
arth 1          algeria 3     rather 5      that 7        her 9
id 2            and 3         than 5        the 7         property 9
and 2           for 3         the 5         only 7        passes 9
i'm 2           her 3         guerrillas 5  next 7        to 9
sure 2          army 3        it 5          of 7          her 9
the 2           protec 3      is 5          kin 7         next 9
cattle's 2      earted 4      p 5           seems 7       of 9
next 2          commitment 4  e 6           to 7          kin 9
of 2            towards 4     a 6           be 7          under 9
kin 2           our 4         narrow 6      a 7           the 9
have 2          kith 4        view 6        cousin 7      intestacy 9
been 2          and 4         of 6          in 7          rules 9
informed 2      kin 4         who 6         droitwich 7   that 9
but 2           overseas 4    is 6          1 8           me 9
```

Figure 5.2
Tokenised Output for 'kin'

Here the numbers simply refer to the concordance line in which each word occurs. Note that every word is converted to lower case, so that different case forms of the same word can form links. If the resulting list is sorted alphabetically, the numbers of the now adjacent identical tokens can be collated. Any word which only occurs once, that is, only has one number in the collated list, is thrown away together with its number, since any such word cannot be forming a part of any recurrent pattern. Once this has been done, what remains is:

```
a 6 7
and 1 2 3 4 5 6 9
been 2 7
for 1 3
her 3 9
in 3 7
is 2 5 6
kin 1 2 3 4 5 6 7 8 9
kith 1 3 4 5 6
next 2 7 9
of 2 3 4 6 7 9
our 4 6
that 2 7 9
the 1 2 5 7 8 9
to 1 6 7 9
```

Figure 5.3
Collated Output for 'kin'

This is the basic information which we require in order to identify lexical cohesion patterns, since it tells us which words occur in which lines. In this instance, the words are allowed to occur anywhere in the line, relative to the node, but this need not be the case, as we shall see in the next chapter. To look at the information for the word 'kith' as an example, the numbers following the word inform us that there is a link between lines (1), (3), (4), (5) and (6), instantiated by the presence of 'kith' in each of these lines.

In **abr,** the word types 'the', 'a' and other very frequent closed-set words were expected to be present in almost every sentence and were not therefore valid candidates for establishing *lexical* links. Since they otherwise contributed nothing but a processing overhead they were removed. In *cohort*, also, such words can be given the same treatment, although, as we shall see subsequently, it is arguable that some grammatical words should be included when concordance lines are being processed, since they may participate in repeated phrases and therefore become valid links. In such cases, this might be regarded as an example of *colligation* (qv Hoey 1997). The contribution of the stopwords to the identification of bonded lines will be covered fully in the following chapter, which describes the various parameters which can be specified to the software.

One other item is also automatically removed and that is the node word, since it occurs in every line and would create an unnecessary link to each member of the concordance. The node and stopwords are thus discarded from the collated list, leaving the remainder shown below:

```
kith 1 3 4 5 6
next 2 7 9
```

Figure 5.4
Collated Output for 'kin' – Stopwords Removed

This list is known as the *wordlist* and the numbers following each word on the list are used to generate the contents of the matrix referred to earlier. Each pair of numbers represents one link, that is, one word in common between the lines in question. The numbers are entered into the matrix a pair at a time, thus 'kith 1 3 4 5 6' would be broken down into all its constituent pairs – (1,3), (1,4), (1,5), (1,6), (3,4), (3,5), (3,6), (4,5), (4,6) and (5,6) – and for each pair the element of the matrix with the corresponding address is incremented. Note that this creates a triangular matrix (as seen in Figure 3.1), since each pair is only entered in the order in which they occur in the wordlist; so for example, the pairs (3,1), (4,1), (5,1) and (6,1) are not used. This means that the same information is not stored twice, which has implications for the way in which the matrix is read later on.

Once all the lines in the wordlist have been processed, the following matrix is created:

```
     1
2    0 2
3    1 0 3
4    1 0 1 4
5    1 0 1 1 5
6    1 0 1 1 1 6
7    0 1 0 0 0 0 7
8    0 0 0 0 0 0 0 8
9    0 1 0 0 0 0 1 0
```

Figure 5.5
Matrix for 'kin' Concordance Lines

The bold numbers down the left-hand side and the diagonal represent the line numbers and the numbers in the cells display a count of the links between the lines. Reading

across from **3** and down from **1** there is the number 1, which means that there is 1 link between lines **1** and **3**. This is the link established by the word 'kith', as it occurred in both lines (1) and (3).

The next stage is to decide how many links must exist between two lines in order for a *bond* to be established. In the example matrix, however, there is no pair of lines which has more than one link. If we are to get any bonds at all, therefore, we must use a threshold of one link per bond. This means that the number of bonds for a line, in this instance, will equal the number of links it has with other lines. Because we did not enter all the reversed pairings into the matrix and are therefore dealing with a triangular matrix, to get the total number of bonds it is necessary to read across and down. So, reading across from, for instance, **5**, we find three links, i.e. three bonds (since one bond here equals one link), and reading down from the **5** on the diagonal we find one further link, that is one bond. This gives line **5** a total score of four bonds.

The bond scores obtained in this way are accumulated in a list external to the matrix and, when the entire matrix has been read, are attached to the original concordance line to which they refer by number. The lines are then sorted by their score, in descending order. Those lines which contain repeated features should by this method be sorted to the top. If a line does not score anything then it is disregarded, that is, it will not be included in the output. Based on the above matrix, the original sample of lines is ranked as follows:

```
4 1  barrier excuses. As for the "kith and kin" appeal, to quote the Reverend Arth
4 3 ted backing of France for her kith and kin in Algeria and for her Army protec
4 4 earted commitment towards our kith and kin overseas." Identity of "race, lang
4 5 f they do lecture their white kith and kin rather than the guerrillas, it is
4 6 e a narrow view of who is our kith and kin. Religion very properly tends to e
2 2 id. "And I'm sure the cattle's next of kin have been informed but is that qui
2 7  been seeing to that. The only next of kin seems to be a cousin in Droitwich.
2 9 and her property passes to her next of kin under the intestacy rules. That me
```

Figure 5.6
Original Concordance Lines with Attached Bond Scores

Here, the first number on each line gives the count of bonds and the second number is the original line number. It should be noted that line 8 did not score anything and has

therefore not been displayed, although this is entirely a presentational issue – lines scoring zero bonds can optionally be retained in the output. The fact that line 8 is not present means that it does not exhibit bonding with any of the other lines. As mentioned previously, this is the only line from the original set of concordances which does not have either of the recurrent 'next of kin' or 'kith and kin' features. Conversely, the lines containing the phrase 'kith and kin' have been placed together in first position, since this is the most frequent pattern, scoring four bonds. They are followed by the lines for 'next of kin', which score two bonds. Naturally, if there had been equal numbers of lines for each feature, the lines for 'next' might have been intermingled with those for 'kith'. If required, however, they could be differentiated either by a simple sorting of the concordance or by more sophisticated methods such as cluster analysis (qv Kaufman & Rousseeuw 1990) using the bond information stored in the matrix to derive similarity measures.

In this configuration of the software, where only one feature per line is being used to identify links, the bond scores can be interpreted as the number of other lines which share the feature in question. Thus the 'kith and kin' lines all score four, since from the point of view of a given line there are *four* other lines which exemplify the same phrase.

5.3. Other Facilities of the Software

5.3.1. Stopword List

As stated earlier, the role of the stopwords in the abridgement software was to exclude the closed-set grammatical items from the wordlist so that only lexical items could qualify as links. It was also mentioned that these higher frequency words might however be of use when concordances were being processed. In order to investigate this possibility, additional stopword lists were created and a mechanism was provided whereby a specific list could be selected and a set of output created on the basis of that list. To accommodate

this, the software automatically includes the name of the stopword list in the filename assigned to the output file. This allows several runs to be made of the program without fear of overwriting the output from the previous run and means that successive runs can be compared more easily. The various stopword lists and the effect of using each one is covered in detail in the following chapter.

5.3.2. Positional Specification

It was noted in an earlier chapter that the concordances used as input to the selection process are formally defined. This allows certain assumptions to be made about their format and introduces the possibility of new link types uniquely applicable to this type of 'text'. To exploit the format of the concordance line, a tailored version of the tokeniser can be used which attaches not only the line number to each token, but also its position relative to the node word, be it a loose before-vs-after distinction or a more precise numerical distance expressed in terms of the number of words separating the word from the node. In the wordlist, this positional information is separated from the word itself by a colon, so that other routines which act upon the wordlist such as the stopword functions (see below) can recognise the characters which comprise the word. In the case of the 'relative' specification, line (1) of the 'kin' concordance

```
barrier excuses. As for the "kith and kin" appeal, to quote the Reverend Arth
```

would be tokenised as:

```
barrier:- 1
excuses:- 1
as:- 1
for:- 1
the:- 1
kith:- 1
and:- 1
appeal:+ 1
to:+ 1
quote:+ 1
the:+ 1
reverend:+ 1
arth:+ 1
```

whereby all words before the node word ('barrier' ... 'and') have a ':–' appended and all the words following the node word ('appeal' ... 'arth') have ':+' added to them. In addition, of course, each word is labeled with the line number from which it came; in this instance the presence of the ' 1' at the end of each line indicates that the words are all from line (1). The positionally-exact tokenisation would be:

```
barrier:-7 1
excuses:-6 1
as:-5 1
for:-4 1
the:-3 1
kith:-2 1
and:-1 1
appeal:1 1
to:2 1
quote:3 1
the:4 1
reverend:5 1
arth:6 1
```

which has the same format as just described, except that the simple +/– positional distinction is replaced with a figure explicitly stating the exact distance between the node word and each token. This is similar in function, and inspired by, the columnar display format of collocates provided by the *picture* software described in Chapter 2. The additional information attached to each token can be used by the next stage of the process. This is described fully in the next chapter.

5.3.3. Link Threshold

In the 'kin' example given above, one link was allowed to generate a bond. Under some circumstances, it is desirable to increase the number of links required and so the Link Threshold parameter used in the abridgement software is retained. Its default value, however, is decreased from three links to just one. The role of this threshold in the selection of concordance lines will be discussed in the next chapter and the way in which it combines with other parameters will be covered in Chapter 7.

5.4. Span Size

An obvious characteristic of concordance lines, but one which sets them apart from the sentences of a normal text, is that they all contain the node word. As well as allowing links to be positionally defined, as we saw above, this facility enables the software to impose a limit on the maximum distance between the node and a link word, to specify, for example, that only words within four words either side of the node are allowed to enter into links. This concept of a restricted context or *span* around the node word is borrowed from collocational analysis, for which four words has often been cited as a usable value (Sinclair 1987b, Collier 1993).

In the software used for this study, the span size is a binary option, in that it is either set, in which case the default span of four words is used, or it is unset, which allows potentially all words in the line to form links. The reason for this parameter and the justification for the choice of four words of context will be put forward in the chapter which follows.

5.5. Conclusion

This concludes the introduction to the facilities of the abridgement software and the concordance line selection software, *cohort*, derived from it. The two subsequent chapters will expand upon the nature of the individual parameters and examine the ways in which they interact.

Chapter 6

Parameters to the Software

6. Parameters to the Software

6.1. Introduction

In the current software system, several parameters, or variables, are involved in the process of calculating bond scores for a set of concordance lines, via the intermediate stages of the matrix and wordlist. This chapter sets out to define the nature and effect of each of the parameters and will explore how the parameters operate in isolation, while the following chapter will cover the interaction of these parameters. Where any parameter differs considerably from those utilised in the abridgement system, this too will be mentioned.

It will be noted that different parameters come into play at different stages in the process. Some affect the creation of the wordlist, while others come into effect later in the process and may therefore influence the contents of the matrix or the assignment of bond scores.

The fact that the various parameters operate at different stages means that altering the value of a single parameter can bring about a substantial change in the nature or extent of the output. The remainder of this chapter will therefore describe how and when each parameter is applied in the process, the exact influence which it exerts independently and, by means of example, the overall effect of modifying it.

6.2. Stopwords

6.2.1. Definition

In the automated abridgement system, links were established on the basis of the repetition of *lexical* items across sentences. In order to limit the forming of links to lexical items only, a *stoplist* was implemented, this being a mechanism whereby a list of word types could be specified which would be excluded by the software from the link analysis. This list consisted largely of closed-set items such as articles, prepositions, modal and auxiliary verbs and proforms. The members of this list were selected on the basis of their non-

lexical nature, but they are also among the most frequently-occurring word types in the language. The justification for their exclusion from the analytical process was therefore twofold. Firstly, they were not lexical in nature and should not properly form part of any profile of the lexical-cohesive characteristics of the target text. Secondly, because of their high frequency, many of these items might be expected to occur with near-random distribution in any given text and thus to contribute links where none would otherwise be found. As far as creating the profile of lexical links within a text is concerned, it is as if the members of the list do not exist at all within the text. The forms which are thus excluded are termed 'stopwords', since they are entirely 'stopped' or excluded from the analytical process.

The concept of stopwords is found in many information retrieval applications. Perhaps the most frequently encountered of these are the World Wide Web search engines (http://www.altavista.com, http://www.excite.com, http://www.searchuk.com), where a search for 'fish and chips' is generally translated into a Boolean query such as 'fish AND and AND chips'; the stopword 'and' is then deleted, resulting in a search for 'fish AND chips' (AND being the logical operator, not the original search term 'and'!). This makes searching the Internet for references to your favourite pop groups 'And And And' and 'The The' nearly impossible. AltaVista, in particular, reports how many occurrences of each of the search words it encountered, and can be seen to flag particular, very frequent words as 'ignored'. The same stopword mechanism is in operation in search applications which run on CDROM databases, examples being Personal Librarian (Independent, Financial Times) and BRS (Northern Echo).

In the concordance line selection process, any item on the stopword list is entirely excluded from the process of link identification. That is to say that as a particular concordance line is examined, any occurrence of an item from the list of stopwords is 'skipped', is not included in the wordlist and from then on plays no further part in the analysis. As far as the content of the stopword list is concerned, the appropriate treatment of

stopwords is not as clear cut as in **abr.** Certainly, many items which appeared in the original stopword lists can be expected to contribute to the identification of links, since they will undoubtedly play an important part in the collocational patterns of particular node words. As an example, imagine trying to identify the important collocates of the word 'ante', when neither 'up' nor 'the' may be included in the analysis because they are on the stopword list. The *cohort* program retains the stopword mechanism, although the stopword list, as will be described below, can be varied in order to tailor the system to the task in hand and to allow for comparison of the effects of various sets of stopwords.

6.2.2. Values and Effect

Seven different lists of stopwords have been employed in the testing procedures. Six of the lists fall into two quite distinct categories, with three lists in each category. The seventh list is quite simple to describe, since it has no members at all.

The category containing the first three lists is defined *functionally*, that is, in a similar fashion to the original stopword list used in the automatic abridgement system; it contains items which are for the most part considered to be grammatical.

Original List This list consists of the items mentioned earlier: articles, prepositions, modal and auxiliary verbs and proforms. It is unaltered from the list used by the textual abridgement software developed for British Telecom and is therefore called **bt.**

There are two major reasons for the use of an alternative stopword list. Firstly, the aims of the analysis are quite different; there seems to be little justification in continuing to employ stopword lists which were identified on the grounds that they helped to produce satisfactory abridgements when the purpose of the exercise is not to create abridgements. Secondly, the input to the current system is no longer naturally-occurring text, but rather concordance lines, which, as seen in the previous chapter, have characteristics of length and cohesiveness very different to those of sentences. Two alternative versions of the

stopword list were therefore explored.

Original List minus Prepositions We know that the collocational behaviour of certain words will tend to 'pull in' items which may have been regarded as stopwords in the textual system. Examples of this would be phrasal verbs which would naturally attract certain particles 'moon *about*' or 'mourn *for*', or prepositional phrases such as '*over* the wall' or '*under* the counter'. If the original list were to be used, then the possibility of these prepositions forming a link would be excluded. It is obvious, however, that in the phrases cited above they are forming part of a recurrent pattern and a new list, devoid of prepositions, was therefore created. Being derived from the original list **bt,** it was named **btb.**

Articles and Pronouns This list is a still smaller subset of the original list and consists solely of articles and pronouns. It is intended to stop only the most frequently-occurring 'noise' words from being included in the analysis, while still being functionally defined. It is referred to as **arts-prons.**

The stopword lists which have so far been described are defined in terms of the type of closed-set items that they contain. In contrast to this, a second set of stopword lists has been defined on the basis of word frequency. These lists, **top50, top100** and **top150,** are made up of the most frequent word types in the BCET corpus, containing fifty, one hundred and one hundred and fifty items respectively.

The purpose of selecting stopword lists on the basis of frequency is to draw away from the purely grammatical definition of the original system. Since the nature of the input has changed so substantially, it might be argued that function of the stopword list should also be modified. Since we are no longer interested solely in lexical items as links, the inclusion of non-lexical items as stopwords can no longer be justified on a functional basis. Just as with natural-language texts, however, the possibility of randomly occurring noise words still exists, and the stopword mechanism is retained in order to compensate for this phenomenon. By eliminating the highest-frequency words from the analysis, simply on

the basis of their frequency and without explicit regard to their grammatical class, the system is better able to accommodate the existence of the 'noise' words in the concordances. The exact level to which these randomly occurring high-frequency words are filtered out can be controlled by the choice of size of stopword list. The inclusion of more items in the list results in a greater suppression of the 'noise', but can also remove valid items from the process if too large a list is used. The words 'explicit regard' in the phrase 'without explicit regard to their grammatical class' are used advisedly, for it is a fact of the language that many of the most common word types in a corpus are indeed closed-set items. Removing these items from the input will thus *implicitly* remove a number of non-lexical items, but as mentioned previously, the effect of this can be monitored by employing a range of frequency-defined stopword lists. If too many links are being detected, that is, a large number of lines are receiving similar bond scores, then a stopword list containing more items can be selected; too few, and a shorter list can be brought into play.

The **zero** stopword list causes all items in the active concordance line (i.e. any item not screened out by any of the other parameters) to become potential participants in the formation of links. In addition to being of interest in its own right, this represents a useful means of determining the exact effects of a given (non-zero) stopword list. This can be achieved by running a set of concordances through the process with a particular stopword list in place and then re-running the same set with the empty stopword list and with all other parameters unchanged. Any change in the contents of the wordlist, the matrix or the output can then be identified and, since no other parameters are modified, attributed to the influence of the stopword list.

Depending on its size and nature, the stopword list will interact with the other parameters in different ways, but on the whole a list containing few or no items will be more likely to allow many links, while a long list of stopwords will probably result in fewer links being created. This supposition will be tested in due course – see Table n(H1.2 for details.

The various stopword lists are summarised in the table below and can be found listed fully in Appendix 1.

Name	Number of items	Number of non-lexical items	Number of lexical items
bt	271	267	4
btb	222	218	4
arts-prons	15	15	0
top150	150	140	10
top100	100	98	2
top50	50	50	0
zero	0	0	0

Table 6.1
Summary of Stopword Features

It is worth noting that some items which might be regarded as lexical have been included in the stopword lists. These tend to be high-frequency, discourse-organising words ('time', 'day' 'people') or homographs in which the non-lexical form is by far the commonest, such as 'can', 'being' or 'mine'.

There is no question that some of the ambiguity of these items could be resolved by the addition of part-of-speech (POS) information, but, as noted in Chapter 1, POS tagged data was not readily available when this study was initiated, nor was any POS-tagged text ever processed using the original abridgement software. POS tagging is certainly something which could be incorporated into the *cohort* software to resolve certain ambiguities and decrease the number of spurious links – this is discussed in more detail in the final chapter.

As noted above, the stopword list directly affects the contents of the wordlist – those words which are candidates for the creation of links – and as such it is the first of the parameters to come into play. This gives the stopword parameter a particular importance and it is worth taking some time to explore fully the impact of varying the stopword list

on the overall results of the software. To illustrate the effect of the various lists, the link and bond analysis was carried out using each list, running on a set of 176 concordance lines for the node word 'exchange'. For each stopword list, the number of members of the list of link words (the *wordlist*) and the total number of links was recorded. As an example, given the following wordlist, the number of link words would be four and the total links eleven, since eleven line numbers are mentioned in the wordlist.

```
commission 2 130
control 35 65
controls 32 75 173
corn 10 16 49 160
```

The results from the various runs of the program are presented in the table below. The number of items in each stopword list is also repeated from the previous table to aid comparison.

Stopword List	No. of Link Words	Total Links	Number of Stopwords
bt	145	422	271
btb	159	571	222
arts-prons	221	1,031	15
top150	128	350	150
top100	147	410	100
top50	188	561	50
zero	236	1,361	0

Table 6.2
Effect of Stopwords on Links

An interesting point to note is that relative to their size, the frequency-based lists restrict the formation of links to a greater degree than the original, word-class-defined ones. The two lists **bt** and **top100** allow approximately the same number of link words (145 vs 147), yet **bt** has nearly three times as many members as **top100**. If a plot is made of the

number of stopwords against the number of link words, another interesting fact emerges:

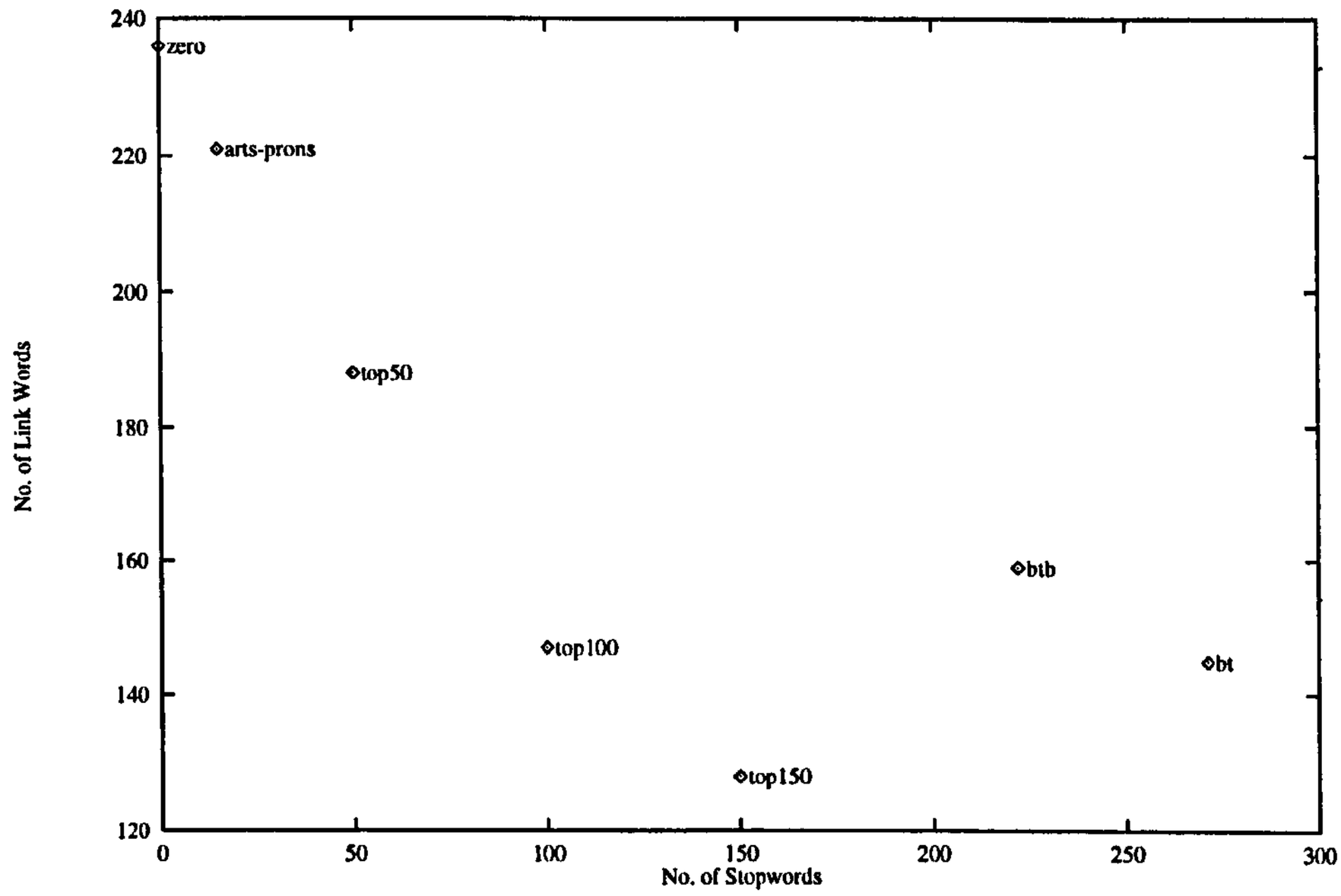

Figure 6.1

Stopword List Size vs Number of Link Words

Notice how the frequency-based lists, as well as **zero**, fall within a regular curve, while **btb** and **bt** stand alone. This is perhaps attributable to the fact that **bt** and **btb** contain a high proportion of relatively low-frequency items, such as enclitic versions of personal pronouns, 'we've', 'he'll' etc, which did not occur in the 'exchange' concordance and will therefore have played no part in limiting the formation of links. The **arts-prons** list fits the curve quite well since, although based on word-class, it is actually a subset of **top50**, that is, it contains no items not found in the frequency-based lists. It does however contain the pronouns 'I' and 'we', which are not present in **bt**.

Moving on to bonds now, the following table shows how many of the total 176 lines achieved bonds using each of the stopword lists and a link threshold of one. The total number of bonds which were formed is also shown:

Stopword List	Number of Bonded Lines	Total Bonds	Number of Stopwords
bt	158	1,226	271
btb	172	4,618	222
arts-prons	176	12,404	15
top150	153	902	150
top100	162	1,090	100
top50	172	1,588	50
zero	176	21,602	0

Table 6.3
Effect of Stopwords on Bonds

The effect represented here is similar to the one observed for link formation. The similarity can be clearly seen by plotting the number of bonded lines against the stopword list size:

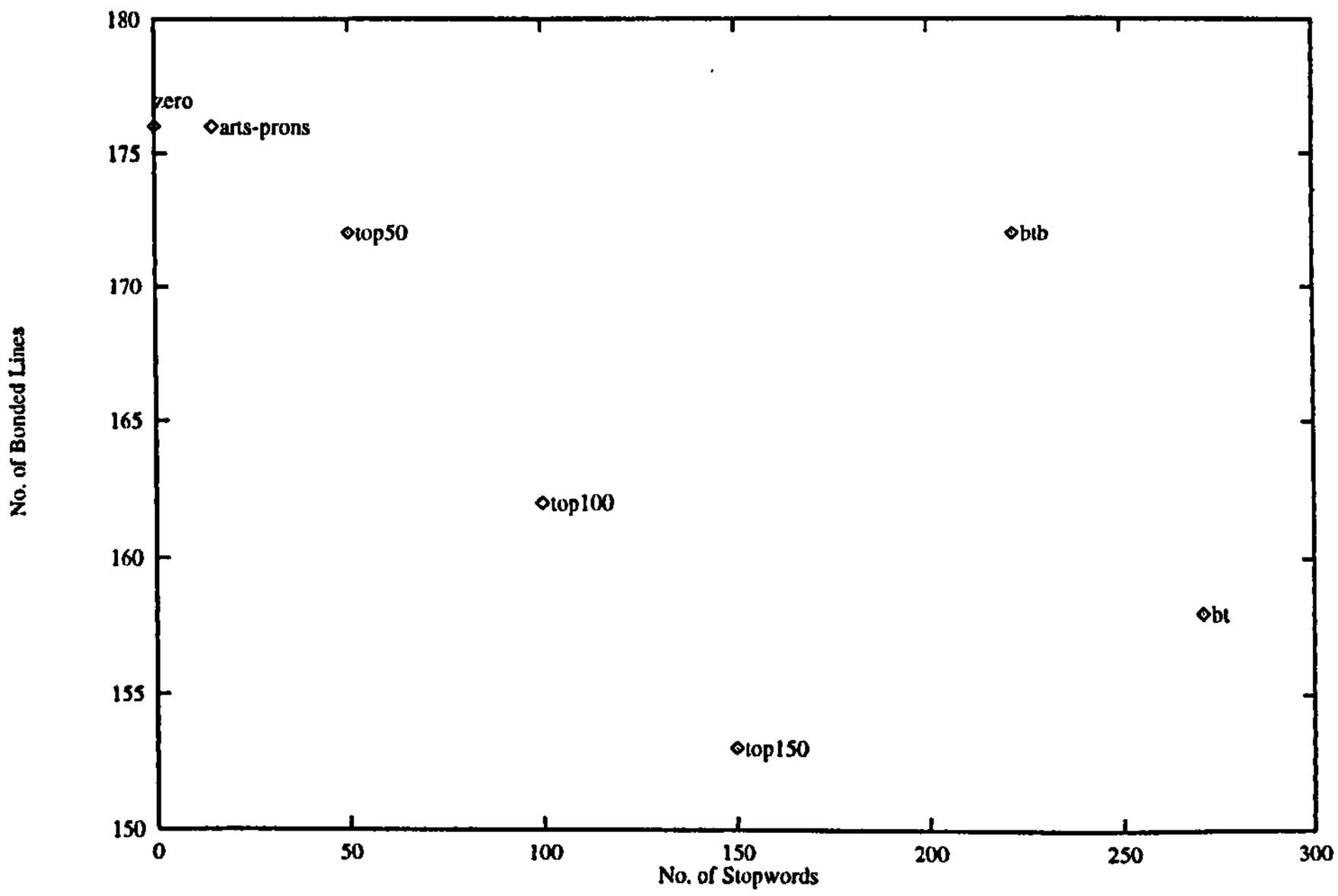

Figure 6.2
Stopword List Size vs Number of Bonded Lines

Again, the frequency-based stopword lists, including **arts-prons**, display a uniform relationship between the number of stopwords and the number of lines which enter into

bonds, while the remaining lists form a separate cluster. Compared with the same points in the previous plot, a slightly shallower slope is presented, since the lowest number of bonds achieved using any stopword list is 153 (for **top150**), a reduction of only 23 or 13% from the maximum 176 achieved using **zero**. The change in link formation described earlier, however, displayed a greater change: from 236 link words (**zero**) down to 128 (**top150**), a reduction of 46%. This would seem to indicate that the increase in the number of stopwords has a somewhat smaller effect on bond creation than on link formation. Interestingly, on this plot **zero** and **arts-prons** both allow *all* the lines to acquire bonds, although the number of bonds received by each line is somewhat higher for the **zero** list.

As regards the total number of bonds allowed by each list, the two shortest lists, **zero** and **arts-prons**, allow several thousand bonds to be formed, whereas the three frequency-based lists cause the greatest reduction in bond formation relative to their size. The fifteen items in **arts-prons** bring about a reduction of 9,198 bonds (21,602 – 12,404), the 50-item list causes a further drop of 10,816 (12,404 – 1,588), yet the 100-item list reduces the total bonds by only 498 and **top150** eliminates only another 188 bonds. It would thus appear that for the frequency-based lists (if one also includes **zero** as a list of the 'top zero' items) a diminishing-returns situation arises. Given the nature of the words on the stopword lists, this is not a surprising result. The word types on the frequency-based lists are so common that they can be expected to eliminate a large percentage of the possible bonds. This can be demonstrated by calculating the total corpus frequency of all the items on each stopword list and then expressing it as a proportion of the total tokens in the corpus. In the table which follows, this calculation has been performed on the basis of the frequencies in the written component of the BCET, from which the frequency-based lists were originally derived. The lists are this time presented in ascending order of size.

List	% Corpus Tokens	No. of Stopwords
zero	0.0	0
arts-prons	18.4	15
top50	41.4	50
top100	49.6	100
top150	53.8	150
btb	43.9	222
bt	50.6	271

Table 6.4
Proportion of Corpus accounted for by Stopword Lists

It is noticeable that firstly, these lists do indeed account for a large proportion of the corpus, but that also there is a fairly rapid levelling off in the percentages at around the 50% level after the first three items in the table. It was noted earlier that relative to their size, the word-class-derived stopword lists, **bt** and **btb**, allowed more bonds to be formed than the frequency-based lists. In the above table the reason for this becomes clear, since the percentages show that **bt** and **btb** do not account for substantially more tokens than any of the other lists, despite the fact that these two lists contain the most stopwords.

This concludes the discussion of the stopword list as an independent parameter. Its interaction with the other parameters of the system will, however, be discussed more fully in Chapter 7.

6.3. Positional Specification

6.3.1. Definition

In the automatic system of textual abridgement, the existence of a link was entirely independent of its position within the sentence, so a lexical item present near the beginning of sentence *S* could form a link with a lexical item at the end of sentence *T*. For the purposes of forming an abridgement, the positioning of the links within the sentence was immaterial, since the text was naturally-occurring language, and no *a priori* assumption

about the relationship among lexical items established as links could be made.

In the case of concordance lines, however, certain expectations can be made of the 'text', since it is known that, firstly, every line contains the node word and that, secondly, the collocational behaviour of the node word will determine that other items – the node's collocates – will be present within the concordance. In addition to the fact that these items are likely to be present, there is also a strong likelihood that their position relative to the node word will be narrowly defined. In order to identify the influence of this phenomenon and, if possible, to exploit it, the Positional Specification parameter is employed. It enables the software to make use of the regularity of form of the concordance line, chiefly that each line is exactly the same size (in characters), contains approximately the same number of tokens and includes the node word in roughly central position. As we shall see in the section on 'Values', this last attribute makes it possible to specify the location of a link relative to the node word. Sentences in a text, however, do not exhibit this kind of consistency, and so this means of positional definition is not possible in either of the abridgement systems.

The Positional Specification must be defined in relation to the consistency with which a given item occurs in the same position relative to the node word within the concordance line. If a strict value is used, then items must occur consistently in the same position in order to register a link. A less strict value implies a looser definition of what constitutes a link. The exact values which can be applied to this variable, and the relative strictness of these values, are described in the following section.

6.3.2. Values

The kinds of feature which can act as a link are more numerous in Hoey's original manual analysis than in the *abr* or *cohort* systems. This is for the most part due to the difficulty of getting a computer to recognise features such as pronominal reference, paraphrase or substitution. Such subtle distinctions, which rely upon human perception or

real-world knowledge, are (as yet) entirely intractable computationally.

Given these limitations, the automatic abridgement system concentrated on just two kinds of repetition: simple and complex (lemmatised) repetition. This basically involved reducing each lexical item to its base (uninflected) form so that, for example, 'planets' would link with 'planet'.

The concordance line system makes use of an even smaller range of the link types identified by Hoey; indeed there is only one type of link which is present in the other two systems, namely that of simple repetition, that is, a word may only link with other occurrences of itself in identical form. This is not to say that this could not be extended to include, say, lemmatised repetition, but for the purposes of this exercise it has been decided to exclude this kind of link, given the large number of possible parameter combinations already present. This decision can be further justified by the argument that in this scenario lemmatisation would be inherently information-losing, since so many of the collocational patterns which we wish to detect involve only a limited subset of a lemma (Renouf 1986). The addition of lemmatisation would also call into question the validity of using a single set of concordance lines, that is, the contexts of a single type, since it might then be argued that one ought to examine the corpus evidence for all parts of the lemma.

Whilst there is only one kind of link common to all three systems, there exist link types which are unique to *cohort* and their use is controlled by means of the Positional Specification parameter. The concept of positionally defining links does not form part of either of the abridgement systems, being entirely novel to the concordance line system. The nearest equivalent to this type of link is to be found in the doctoral research of Peng Wangheng at the University of Liverpool (Wangheng 1998), who is investigating the positioning of the links within the sentences of a text with a view to establishing the correspondence between link position and the theme-rheme constituents of the sentence.

The Positional Specification parameter may currently take one of three values, which vary in the strictness of their definition of what constitutes a link. These are defined as follows:

Raw

This is the least strict and implies that *no* limitations exist on where in the concordance line an item must occur. It is therefore the closest equivalent of Hoey's Simple Repetition class of link, although it must be remembered that, depending on the stopword list in operation, it is possible for non-lexical as well as lexical items to form links. This is a significant divergence from the definition of links in either of the abridgement systems and is exemplified by the two lines for the node word 'exchange':

```
(1) ate a small plot on the worst land, in exchange for agricultural and even dome
(2) r thought of applying for the Euphoria exchange: "Not really, Gordon. It would
```

Assuming an empty stopword list is in use, the item 'for' in (1) would link with the same item 'for' in (2). The usefulness or otherwise of this is discussed in the 'Effects' section which follows.

Absolute

This is the strictest level of positional specification. At this setting, an item must occur in exactly the same position relative to the node word in two lines in order for it to form a link. In lines (3)–(7) an instance of this can be seen, again using 'exchange' as the node. Here the item 'in' at position **–1** (one to the left) would be identified as a link between the three lines (3)–(5). This would not be the case for line (6), where 'in' occurs at position **–2** and is therefore not linked with the instances of 'in' where it occurs in the **–1** slot. Similarly, 'for' in the **+1** (one to the right) position is available to form a link in lines (3)–(5), but the 'for' in line (7) is not allowed to link with these. The presence of 'in' or 'for' in any other position, then, is not regarded as a potential link.

(3) e disposed of, and offered for sale in exchange for cash - and when cash is no
(4) , price controls and food subsidies in exchange for voluntary wage restraint,
(5) hat the offence might be overlooked in exchange for a consideration: they woul
(6) imise our clients' exposure in foreign exchange. We tell them what's happening
(7) for how they would live. She would not exchange her solitude for anything. Nev

Relative

This value implements an intermediate level of rigidity between the 'raw' and 'absolute' link types. It allows the linked item to 'float' slightly, decreeing a link to have occurred between two lines if an item is repeated on the same side (left or right) of the node word in the two lines. This is exemplified in lines (8)-(10) below, where a link can be established between all the occurrences of the item 'ideas', even though it occurs one to the right of the node in (8), three to the right in (9) and two to the right in (10). The important feature is that 'ideas' occurs in the right-hand context in each case.

(8) and see me again. It'll do us good to exchange ideas." She could have been gl
(9) came from Berlin and abroad, eager to exchange the new ideas that were racing
(10) d enjoyment. The justification is the exchange of ideas, and the value of thi

6.3.3. Effect

The relationship between Positional Specification and the presence of collocational patterning in the concordance lines has already been mentioned. The effect of setting this parameter at any value other than 'raw' is to accentuate those lines which exhibit the collocational characteristics most frequently repeated in other lines. If, for example, lines (3)-(5) were to be processed with a Positional Specification of 'raw', then they would be merged in with all the other lines which contained the items 'in' and 'for' *anywhere* in the line. They would, therefore, receive a link score on the basis of the presence of these two items, but so would all the other lines containing 'in' and 'for', regardless of the positioning of these items relative to the node word. If a Positional Specification of 'absolute' is employed, however, all the lines in which 'in' and 'for' appear 'randomly', i.e. not forming part of a repeated positionally-constant collocational pattern, are filtered out, since they fail to achieve any links. Those lines such as (3)-(5) however, because they exhibit

consistency in the position of the collocates, do receive links.

The type of collocational patterns which can be isolated in this fashion will depend on the particular value of the parameter. This relates to the 'fixedness' of the collocates in relationship to the node word.

Collocational characteristics consisting of items which regularly occur at the same location will be better identified by using a value of 'absolute', viz the 'in exchange for' examples shown above.

Collocational patterns whose individual constituents are more free to drift, on the other hand, will benefit from processing with this parameter set at the intermediate value, 'relative', as can be seen in lines (8)-(10). In the examples given, 'ideas' is identified as a link, even though it occurs across a range of positions in the right-hand context of the node word. Any occurrence of this item in the left-hand context, however, is rejected as a link. The set of links detected using the 'absolute' setting will naturally always be a subset of the links identified using the 'relative' setting, but using the less strict value will cause the selected lines to be 'watered down' by the addition of lines exemplifying patterns which are more variable.

Setting this parameter to 'raw' will tend to identify patterns which are perhaps too variable to be detected using one of the stricter settings. This might be the case if one were analysing the concordances of a very common type, a preposition, for example. In this instance it might be expected that the node word would be taking part in a large number of collocational patterns, some of which may be quite loosely defined positionally. If we take a particular phrasal verb particle, 'up', as an instance of this, it can be shown that links would be lost by using a strict value for this parameter. Supposing that, for example, lines containing the sequence 'put **up** prices' are present in a set of concordance lines exemplifying the node 'up', then we would sensibly want these to link to lines which contained the sequence 'put prices **up**'. 'Put', though, would fail to form a link using the absolute setting and 'prices' would be excluded from linkage using either of the other

settings, since it is free to 'swap sides' around the node word 'up'. The 'raw' value, however, permits both 'put' and 'prices' to form links.

6.4. Link Threshold

6.4.1. Definition

To recap briefly on the terminology of textual abridgement, a sentence *S* is said to have acquired a *bond* with another sentence *T* if the two sentences have a specified number of lexical items in common. Each of these lexical items common to *S* and *T* is referred to as a *link*. Thus if three lexical items occur both in *S* and *T*, then it can be said that three links exist between *S* and *T*, or, seen from the perspective of the individual sentences, that each sentence has three links. What the *link threshold* does is to define the number of links that a sentence must have in order to acquire a *bond*. This means that if, for example, the link threshold is set at three, and sentences *S* and *T* are related by three (or more) links, then those sentences can be said to be *bonded*. As described earlier, this link information is stored in a *matrix* which lists the number of links between all possible pairs of sentences. Once the matrix is built, a *bond score* can be established for each sentence according to the number of sentences with which it has a sufficient number of links. If the link threshold is still three and sentence *S* has three links with *T*, four links with *U* and three links with *V*, *S* can be said to have three bonds. Notice that even if four links are present, only one bond is established; the same two sentences are never bonded to each other more than once.

In a set of concordance lines, the relationship between links and bonds is identical to that just described: the matrix is first constructed and then the bonds are analysed on the basis of the prevailing link threshold. The definition of the link, as seen in the previous section, is substantially modified however, as is the basic element of analysis, of course. In exactly the same way as for textual abridgement, each concordance line (the equivalent

element to the sentence in textual abridgement) is examined and the number of items it shares with other lines is calculated. This data is stored in the matrix and then, once the matrix is complete, i.e. all the possible pairs of lines have been compared, each line can be assigned a bond score. Any line which receives a bond score greater than zero is said to be *selected*, that is, it will be presented to the user of the software as preferable to any lines which do not acquire any bonds.

6.4.2. Values

In constructing abridgements of natural-language texts, Hoey's manual system and the automatic system commonly used a link threshold of **3** as a starting point. Where the abridgement was required to be shorter, this was raised to **4**, making it correspondingly harder for a sentence to acquire enough links to achieve a bond. The reasons for the choice of these values was discussed in some detail in Section 3.3. For textual abridgement, a further mechanism, the *bond threshold,* came into effect. This restricted the sentences included in the abridgement to those which had acquired the requisite number of *bonds*. Setting a higher bond threshold generally resulted in fewer sentences attaining it, thereby reducing the length of the abridgement.

When using the system on concordance lines, only the *link* threshold is specified when the process is started. This contrasts with the textual abridgement procedure, where an explicit *bond* threshold needs to be set. This is because the ultimate aim of *cohort* is to calculate a bond score for *every* line. Since the lines are presented in descending order of bond score, it could be said that it is the *user* who implements the bond threshold, albeit as a mental process, since there is a point in the bond score range at which the user will decide that lines any lower in the ranking are unlikely to be usable. In order to reduce slightly the number of lines selected, the software can, in fact, suppress any line which fails to acquire any bonds whatsoever. This is the nearest approximation to a bond threshold in the system.

The values used for the link threshold in the concordance line selection system have so far been in the range of **1** to **6** inclusive; any value exceeding this has, for the words examined to date, failed to select any lines at all. This differs somewhat from the values used in textual abridgement, where **3** was regarded as a minimum 'safe' value, that is, the value below which the risk of generating spurious bonds rose sharply. The reasons for this distinction are twofold: firstly, in the way in which a link is defined in this system and secondly in the special nature of the concordances themselves.

In discussing the construction of the wordlist it was mentioned that each line of the concordance contains at least one link to all others in the form of the node word. On this basis alone, it would be safe to lower the threshold to two (since the node is never counted in the link analysis). The influence of collocational patterning must also be considered. A threshold of two implies that there must be at least two repeated items in the environment of the node word, but if, for example, a stopword list is being used which disallows closed-set, non-lexical items as links, this is still quite a strict criterion. The link threshold for the analysis of concordance data is therefore allowed to be as low as one.

In Chapter 4 it was noted that the order of presentation of the lines does not need to be preserved. *All* the concordance lines in the input to the software can therefore be included in the output, but they are re-ordered to reflect the number of bonds which they attract. This being the case, it is less harmful to employ a low link threshold, since the lines will be further filtered when they are sorted, presenting the highest-scoring lines first. In the case of textual abridgement, conversely, weakly linked sentences have to be excluded completely, as they will otherwise intrude on the key sentences; they cannot be simply re-ordered to the bottom of the text.

As to the maximum possible values for the link threshold parameter, the only theoretical limit for textual abridgement is that the number of links cannot be higher than the number of tokens in the sentence. As mentioned earlier, however, sentences can contain large

numbers of words. Despite this, it has never been necessary to raise the link threshold higher than six in order to create an abridgement, although using a higher value would probably still result in an abridged version of the text in some cases. In the case of concordances, the maximum depends to a great extent on the stopword list and type of link being used. Experience with several concordance sets has shown that if only lexical items are being considered as candidates for linking and a strict (i.e. relative or absolute) link type is used, then only duplicate lines or lines containing very fixed phrases acquire any bonds. Given a balanced combination of the other parameters, the link threshold should probably not be set any higher than three, but where few or no stopwords are in action, this has been raised as high as six.

6.4.3. Effect

The end effect of the link threshold is to limit the number of lines which are included in the output from the system. The mechanism by which this is achieved is quite simple, but the effect is not completely predictable, in as much as the relationship between the threshold and the number of lines selected does not adhere to any rule other than the general observation that a higher threshold usually results in the selection of fewer lines.

When a concordance set is being processed, the link threshold is initially set at **1**. At this setting, just a single shared feature can establish a bond between two sentences. This value therefore decrees that a line will acquire a bond for each line with which it shares one or more links. When this threshold value of **1** is used, a large proportion of the lines will receive a bond score, since the probability that each line has at least one link to another line is quite high. If the threshold is increased to **2**, then it is likely, although not certain, that the number of lines which receive a bond score and are thereby selected for output will be reduced. The uncertainty stems from the possibility that *all* the lines selected for output when the threshold is **1** in fact have *two* or more links with at least one other line. This can occasionally be the case when a small set of concordances is used,

the node word of which forms part of an idiom or other fixed phrase, or if a small or empty stopword list is being used.

Naturally, if the threshold is raised even higher, then the likelihood of a line acquiring a bond is reduced further still. The exact nature of the lines that are selected depends to a great extent on the way in which the link is defined, and is thus heavily influenced by the choice of stopword list and by the positional specification which is in effect.

As an example, let us return to the 'exchange' concordance. In the table which follows are the results of analysing the bonds created using a range of different link thresholds with the 'raw' link type and the 'zero' stopword list.

Link Threshold	No. of Bonded Lines
1	176
2	176
3	158
4	108
5	22
6	6
7	2

Table 6.5
Effect of Link Threshold on Bond Formation

A number of points are raised by these results. Firstly, the original threshold selected all of the 176 lines in the concordance and increasing the link threshold from one to two has no effect. All the lines therefore have at least two links. Secondly, for this concordance raising the threshold above six produces no linguistically useful output. The two lines selected at a threshold of seven are in fact near-duplicates:

```
(124) ng deposits). <P 8>  As with all Stock Exchange investment price can go down
(156) t bearing deposits). As with all Stock Exchange investment prices can go down
```

Numbers in parentheses refer to the original line number in the concordance.

The lines selected at the six link threshold, with the exception of the two lines above which are of course also included, are of some interest in that they exemplify fixed phrases:

```
(43)  W  YORK - Prices on the New York Stock Exchange staged a  blue- chip  rally  F
(105) ly escalation into a strategic nuclear exchange between the Soviet Unon and t
(137) ous abbreviation on the New York Stock Exchange. A stock listed as Spud appare
(168) ut escalation into a strategic nuclear exchange. On the other hand, an observe
```

These lines each receive one bond each, due to the fact that they form pairs of bonded lines, linked by the words 'a', 'New', 'on', 'Stock', 'the' and 'York' in the case of lines (43) and (137) and by 'a', 'escalation', 'into', 'nuclear', 'strategic' and 'the' for (105) and (168).

When the threshold of five links is used, the 22 lines selected are:

```
 92 3 he flat and go to a hotel. The rate of exchange in Denmark is heavily against
112 2 means of production, distribution, and exchange is profitability; that any dep
113 2 means of production, distribution, and exchange", meant something quite differ
114 2 means of production, distribution, and exchange". The prose style of the notor
 54 2 and manufactures to the third world in exchange for raw materials and food, is
 46 1 a our Unit Trusts then we have a Share Exchange Scheme whereby you can obtain
 69 1 der our range of Unit Trusts. (( Share Exchange. )) If you already own some st
124 1 ng deposits). <P 8>  As with all Stock Exchange investment price can go down
156 1 t bearing deposits). As with all Stock Exchange investment prices can go down
 43 1 W  YORK - Prices on the New York Stock Exchange staged a  blue- chip  rally  F
137 1 ous abbreviation on the New York Stock Exchange. A stock listed as Spud appare
  7 1  endeavoured to heal the wounds. In an exchange of letters with Mansholt he de
117 1 mpete if we are to earn enough foreign exchange to buy the primary goods we st
165 1 tries will fail to earn enough foreign exchange to maintain our primary base t
 71 1 ds over her identity to her husband in exchange for a small portion of his, sh
 52 1 amental right to adequate treatment in exchange for being deprived of his libe
107 1 man gives food, care and protection in exchange for the different services the
105 1 ly escalation into a strategic nuclear exchange between the Soviet Unon and t
168 1 ut escalation into a strategic nuclear exchange. On the other hand, an observe
 90 1 hat is, to establish the exact rate of exchange at which mechanical energy is
142 1 pproached the defense table, hoping to exchange a few words with them. The  gu
161 1 the night officer and the sister would exchange a few words with us. In my fir
```

The first figure on each concordance line gives the line number and the second the number of bonds acquired. In this set of output, lines with the same bond score have been sorted on the first word to the left. This set of output highlights an important issue. It will be noted that there is little variation in the number of bonds: only one line achieves three

bonds, four get two bonds, and the remainder only one. This is a result of the high link threshold, as it allows few lines to acquire any bonds – in addition to those shown here there are 154 lines which received no bonds at all – and those lines which do enter into bonds do not score highly. Since the aim of the system is to rank the lines, this is not a desirable state of affairs; it would be preferable to have many lines acquiring bonds which should give a larger range of scores and allow a more fine-grained ranking of the lines. When selecting a link threshold, then, this issue must also be taken into consideration.

6.5. Span Size

6.5.1. Definition

Some use has already been made in this chapter of the word *context*, although no formal definition has yet been put forward for the term. It is the function of the Span Size parameter to define the context. The *span* is a term borrowed from collocational analysis and refers to the number of items around the node word which can be regarded as forming part of its environment. Seen in collocational terms, this defines the size of a 'window' onto the items which regularly occur with the node word. The implication of this is that all the significant collocational patterns associated with a particular node word will occur within that window, the corollary of which is that *only* items falling within the window, whatever its size, will be recognised in the course of the collocational analysis. The optimum size of this window, or context, has been the subject of some debate.

6.5.2. Values

The primary motivation for the inclusion of this parameter has been to allow a distinction to be made between a fixed span and a span which encompasses the entire concordance line, referred to here as the 'open' span. For the fixed span, four words either side of the node word were used. This was chosen on the basis of several years of collocational

analysis which have shown that a large proportion, although by no means all, of the collocational behaviour of a given node word can be identified within four words either side. This accords with the findings of other researchers, such as Sinclair (1987b). This span size also offers a reasonable trade-off between accuracy and processing overheads. The use of the open value is mandated by the fact that the raw input to the analysis is in the form of concordance lines and it is this Keyword in Context (KWIC) format which is familiar to many corpus researchers. If a comparison is to be made between the evaluation of the concordances by human beings and the analysis provided by the automatic system, then the open span needs to be an option in the specification of the parameters to the system.

Two major factors influence the choice of span size: the desire to capture a comprehensive picture of the collocational behaviour of the node word and the computational burden of recording the contents of the window. As might be expected, these two factors tend to come into conflict with each other. By setting the span size to a small value, recurrent items may well be missed from the analysis, since there is a greater risk that they will fall outside the span, resulting in an incomplete account of the node word's collocational behaviour or 'profile'. The smaller the span, the more distorted the picture. If the span is made larger, then more items can be included, increasing the likelihood that the analysis will be a comprehensive one. At the same time, however, there is an increase in the overhead, be it manual or computational, of processing the items encompassed by the span.

In the current system, the Span Size is specified in terms of the number of items to the left and to the right of the node which will be considered for inclusion in the process of identifying links. Thus, setting the Span Size as **4** will allow only items which occupy the four slots either side of the node word to form links; any items which occur more than four slots away from the node are excluded. Expressed otherwise, this is equivalent to a context of eight words and, since the same number of words are examined before and after the node, is sometimes referred to as a ±4 span. While it is possible to conduct

collocational analysis on the basis of unbalanced spans, two words to the left and six to the right for example, this has not been pursued in this study, since this type of analysis tends to be slanted towards the investigation of particular word classes, such as phrasal verbs, and it is the goal of this study to cover only those facets of the *cohort* program which are as generic as possible.

A further reason for the choice of four words as the fixed span size is the fact that the majority of concordance lines can be expected to provide at least this much context. As noted in the chapter where we compared the characteristics of sentences and concordance lines, the most likely length of the fixed width concordance lines analysed for this study is between twelve and seventeen words, with 93% of lines falling within this range. Given that one of the words is the node word, a span of anything greater than five is likely to be unapplicable to an increasingly large proportion of lines, while anything over eight will probably only be suitable for a small minority of lines. This is compounded by the fact that in an 80-character concordance, as used in this study and discussed in Section 4.2.2., the node word starts at character 40, so the amount of available left context tends to be slightly larger than the right, viz:

```
(62) blushing, as the following therapeutic exchange demonstrates: Therapist: "Why
```

which has eight words of context available, but five of them are on the left of the node and only three on the right. If a span of ±4 were to be applied to this line, no collocate information would be available from the +4 slot, as no word exists four slots to the right of the node; this would in effect cause a –4/+3 span to be used, which could compromise any statistical analyses one might wish to perform. Obviously the likelihood of a mis-match between the desired span and the amount of context actually available in the line will increase as the span size is raised, which further militates against attempting to extract spans larger than ±4 from the concordance lines. In the next section, the effect of various span sizes will be examined. The main focus will be on the variation in link and bond formation brought about by different spans, but an analysis is also included of the

applicability, in terms of the amount of available context, of each of the span sizes to a trial set of concordance lines.

6.5.3. Effect

The effect of setting a fixed span is to limit the number of items which can be included in the process of link identification. For the 'open' span *all* items present in the concordance line are candidates for linkage with other lines, whereas if a fixed span is employed then any items occurring outside the specified span will be ignored.

In a similar way to Positional Specification, this parameter is closely bound up with the type of collocational behaviour which it will identify. If one accepts the view that a ±4 span is sufficient to encompass the significant collocational patterns of a particular node, then the use of a fixed span of this size should not be a cause of concern that relevant items may be excluded from the analysis. If, on the other hand, such a span is regarded as insufficient in its scope, then the use of the entire concordance line will provide a form of control by which the inadequacies of the fixed-size span may be identified.

In order to investigate the effect of span size more closely, a comparative analysis was carried out, examining the effect on link and bond formation of using spans of various sizes. Fixed-size spans from one to eight words either side of the node were tried, as well as the open span, and the other parameters were:

node word: 'exchange' (176 lines)
stopword list: 'zero'
positional specification: 'raw'
link threshold: 1

The results are given in the table below, which shows figures for the same metrics (link words, total links, bonded lines and total bonds) that were applied in the stopword list section above.

Span Size	Number of Link Words	Total No. of Links	Number of Bonded Lines	Total No. of Bonds
1	44	259	165	2,816
2	78	456	174	7,780
3	117	664	176	11,386
4	154	883	176	15,556
5	190	1,075	176	18,682
6	212	1,224	176	20,738
7	229	1,320	176	21,358
8	236	1,356	176	21,474
open	236	1,361	176	21,602

Table 6.6
Effect of Span Size on Link and Bond Formation

The most striking feature of this table is that all but the smallest two span sizes result in all the concordance lines being bonded (Number of Bonded Lines = 176). This is no doubt attributable to the looseness of the other parameters and will be addressed in a later chapter on the interaction of parameters. There is, however, a good range of variation in the other columns, certainly enough to indicate that the ±4 span is far from comprehensive in its coverage of potential repeated features. In terms of the total number of links, if one takes the 'open' value as 100%, the span of four achieves only 65%. Looking at the total number of bonds, it does slightly better, accounting for 72% of the possible bonds represented by the 'open' total.

It is clear from the above table that the larger spans are causing more links and bonds to be created. As to the value of the extra links established by increasing the span size, however, few strongly recurrent patterns seem to be identified by the addition of the extra context. The qualitative effect of the span is investigated more thoroughly in the evaluation excercise which is described later. A quantative analysis is presented here, however, in the form of a list of the additional link words which are gained by increasing the span from four to open. The figures in the second column refer to the original line numbers in

the 'exchange' concordance, which can be found in Appendix 2, while the final column records whether the words have formed an intuitively valid link. The entries in the last column other than simply 'yes' or 'no' are defined as:

truncated

The link word occurred at the beginning or end of the concordance and was thereby truncated.

marginal

The link word could be considered to be valid under certain circumstances, but not as a cohesive item. This category is explained in greater detail for individual words after the list below.

duplicate

The link word occurred in a context which appears more than once in the corpus, resulting in a duplicate concordance line. This can be attributed to lines which have only minor typographical differences between them, since completely identical lines are filtered out by the software.

Link Word	Line Numbers	Acceptable as Link
al	48 49	truncated
already	12 69	no
also	28 121	no
another	154 173	no
b	64 171	truncated
back	93 125	no
become	110 151	no
been	4 22 57	no
berlin	5 85	no
best	15 23 77	no
britain	2 12	no
came	1 5 13	no
chairman	2 131	marginal
changed	64 132	no
com	88 140	truncated
countries	9 21	no
currency	65 148	marginal

Link Word	Line Numbers	Acceptable as Link
d	66 67 68 70	truncated
decent	30 94	no
deposits	124 156	duplicate
did	42 50 101	no
down	50 124 156	duplicate
e	72 73 74	truncated
each	49 61	no
er	35 80 133	no
escalation	105 168	yes
exchanges	8 36	marginal
f	13 43 81 115	truncated
food	32 54 107	marginal
found	10 131	no
full	6 170	no
give	39 129	no
hadn't	30 31	no
hand	125 168	no
hat	89 90 91	truncated
his	52 71 96	no
home	12 23	no
it's	88 126	no
know	15 133	no
land	60 141	marginal
like	30 78	no
ly	105 106	truncated
maintain	53 165	no
market	63 76	yes
means	112 113 114	yes
miss	10 49	no
n	40 118 119 120	no
nev	22 83	truncated
ng	124 125 126	truncated
out	123 154	no
over	71 172	no
papers	81 140	no
people	41 174 176	no
price	32 124	no
prices	43 156	yes
raw	54 67	no
re	10 17 145 146	truncated
recently	130 170	no
s	93 153	truncated
see	4 33 174	no

Link Word	Line Numbers	Acceptable as Link
since	151 176	no
some	22 69 132	no
st	69 110 117	truncated
t	9 67 105 127 156 157 160 165 166	truncated
th	16 143	truncated
then	46 162 176	no
thought	132 144	no
told	33 85	no
trip	115 171	no
unit	46 69	yes
up	39 48 70 111	no
us	4 50 161	no
value	18 97	no
very	11 63	no
wa	31 169	truncated
well	81 93	no
who	171 174	no
will	70 165	no
wo	82 103	truncated
y	173 174	truncated
yes	35 93	no
your	3 109 141	no

Of the 82 total link words in the above list, 52 are definitely not valid links while a further eighteen are truncated because they occurred at the beginning or end of the line and so have to be ignored. The remaining twelve vary in the degree to which they could be said to be linked; some are obvious collocates, but others have a more oblique connection to the node word:

chairman These lines exemplify noun phrases of the form 'chairman of *some institution*' and so are acceptable as links:

```
(2)    Ruder, chairman of the Securities and Exchange Commission, said Britain, the
(131) olas Goodison, ? Chairman of the Stock Exchange, was asked if he found the lar
```

currency There is certainly a real-world connection between 'exchange' and 'currency' and so the link is a valid one. No paradigmatic or syntactic regularity is present however:

```
(65)  currency accounts was established when Exchange Control Regulations were lifte
(148) riday. You cannot cash a bank draft or exchange foreign currency when the bank
```

down (50)

Line 50 does not fit at all and lines (124) and (156) are linked through duplication. Definitely spurious, therefore:

```
(50)   all of us if we did not calm down. Our exchange was heated. Within a matter of
(124) ing deposits). <P 8>  As with all Stock Exchange investment price can go down
(156)  t bearing deposits). As with all Stock Exchange investment prices can go down
```

escalation This forms part of a long repeated string which does not appear to be attributable to duplication and so is acceptable as a link:

```
(105) ly escalation into a strategic nuclear exchange between the Soviet Unon and t
(168) ut escalation into a strategic nuclear exchange. On the other hand, an observe
```

exchanges This item, like the node word, changes its word class and sense in these examples, so this seems an implausible link:

```
(8)   exchanges; genes within bacteria can  exchange. But, in the past at least, it
(36) Copies were burned on the London Stock Exchange and destroyed at exchanges in
```

food No pattern is observable here other than, perhaps, the fact that food is something which may be exchanged for other commodities, making it a 'real-world' collocate rather than a linguistic one. It nevertheless has some value as a link:

```
(32)  , price controls and food subsidies in exchange for voluntary wage restraint,
(54)  and manufactures to the third world in exchange for raw materials and food, is
(107) man gives food, care and protection in exchange for the different services the
```

land This is a similar case to 'food' in that it is something which is used as a means of exchange:

```
(60)  ate a small plot on the worst land, in exchange for agricultural and even dome
(141) pawn your land for five years or so in exchange for the cash. The moneylender
```

market This has strong associations with trade and exchange and so might be a useful link to include:

```
(63) ce economies (where very little market exchange takes place), this form is on
(76) ee market, for which you need an equal exchange between equal parties. Even wh
```

means is used here as the head of a long noun phrase and seems acceptable as a link. The closeness in value of the line numbers is, however, almost certainly a sign that these citations are drawn from the same source. An experienced corpus user would therefore be wary of making generalisations on the basis of these lines:

(112) means of production, distribution, and exchange is profitability; that any dep
(113) means of production, distribution, and exchange", meant something quite differ
(114) means of production, distribution, and exchange". The prose style of the notor

prices Both occurrences refer to the prices of shares on a stock exchange, so a definite link:

(43) W YORK - Prices on the New York Stock Exchange staged a blue- chip rally F
(156) t bearing deposits). As with all Stock Exchange investment prices can go down

unit This forms part of the nominal group 'Unit Trusts', which is closely connected with share exchange. Incidentally, 'Trusts' occurs elsewhere within a ±4 span, but without 'Unit' and so is not identified as an additional link here.

(46) a our Unit Trusts then we have a Share Exchange Scheme whereby you can obtain
(69) der our range of Unit Trusts. ((Share Exchange.)) If you already own some st

The three words now identified as spurious links – 'currency', 'down' and 'exchanges' – plus the original 52 give a total of 55 spurious links. If the 18 truncated words are ignored, this represents 86% of the additional links created by raising the span size from four to open.

The next analysis is based on using a span of five and compares the results obtained from using a span of four.

Link Word	Line Numbers	Acceptable as Link
already	12 69	no
back	93 125	no
become	110 151	no
been	4 22 57	no
berlin	5 85	no
came	1 13	no
chairman	2 131	marginal
countries	9 21	no
currency	65 148	marginal
did	42 50	no
down	50 124 156	duplicate
er	35 133	no
escalation	105 168	yes
exchanges	8 36	marginal
food	32 54 107	marginal
like	30 78	no

Link Word	Line Numbers	Acceptable as Link
means	112 113 114	yes
miss	10 49	no
people	41 174	no
raw	54 67	no
recently	130 170	no
since	151 176	no
some	22 69 132	no
then	46 176	no
told	33 85	no
trip	115 171	no
up	48 70	no
us	4 161	no
value	18 97	no
very	11 63	no
well	81 93	no

Twenty four (77%) of the thirty one† items are definitely not acceptable as links and of the remaining marginal links examined earlier another two can be rejected, as we have seen, giving a total of 26 spurious links, or 84% of the total 31 words. If we compare the results just presented with those given for the open context, we find that the words 'land', 'market', 'prices' and 'unit' are now excluded.

Having looked at the effects of *increasing* the fixed span size from four words, it makes sense to carry out a contrastive analysis of the effect of *reducing* the context in which links must occur. In the table which follows a list is presented of those link words which would be lost if the span were to be reduced from ±4 to ±3 words.

Link Word	Line Numbers	Acceptable As Link
about	33 51	no
again	20 120 135	no
any	30 86 88 112	no
bank	148 158	yes
because	56 162	no
children	1 44	no
even	60 74 76	no
genes	8 89	yes

† Truncated words were excluded from this list, hence the mismatch between the length of the list (31) and the number of items expected on the basis of the Table 6.6 (36).

given	94 171	no
go	124 156	no
goods	64 145	yes
gray	10 49	yes
had	96 127	no
how	47 75	no
into	105 168	yes
it	14 21 26 121 144 153	no
london	27 36	no
man	22 80	no
materials	54 67	yes
most	122 129	no
old	40 79 101	no
own	69 108	no
primary	117 165	yes
public	88 157	no
received	170 171	no
risk	47 143	no
services	64 107	yes
severe	116 118	no
takes	6 63	no
them	74 98	no
things	18 174	no
through	101 110	no
together	119 150	no
various	53 145	no
where	10 19 63 157	no

For the ±3 span, 27 of the 35 links which would be lost compared to the ±4 span are actually spurious. This corresponds to 77% of the total.

If an even smaller span is used, it can be expected that more links, some of which will be spurious, will be lost. The following list, based on a ±2 span, shows that this is indeed the case. Items which are also in the previous list are suppressed.

Link Word	Line Numbers	Acceptable As Link
bacteria	8 89	yes
by	77 79 111 154	no
cash	73 141	yes
could	4 22 31 164	yes
don't	89 126	no
earn	117 165	yes
floor	119 167	yes

from	55 89 157 164 171	no
has	12 57 106 115	no
have	4 12 46 149	no
her	71 83	no
him	45 115	no
if	69 131	no
little	63 80	no
me	33 77 78 154	no
meet	82 150	yes
money	80 145 158	yes
need	67 76	yes
new	5 43 137	yes
often	6 9	no
one	12 16 64 86 149	no
only	61 138 159	no
other	21 82 168	no
place	15 63	no
production	112 113 114	yes
regulations	35 65	yes
said	2 24 120	no
she	4 83	no
should	12 58	no
small	71 116	no
soviet	105 14	yes
their	82 84 122 125 155	no
vows	79 95	yes
way	22 133	no
were	13 34 65 103 169 171	no
why	58 62	no
within	8 50 53	no
words	22 142 161	yes
x	12 26	yes

Here, 24 out of the 39 items, 62%, are false links. If this result is combined with the comparison of the four-word and three-word spans, then the overall number of links lost by reducing the span from four to two is 74, of which 51 (69%) are spurious.

Although very little evidence is available, there does seem to be a relationship between the span size and the number of valid links which are gained as the context is increased. This is summarised in the table below, which indicates that if the span is increased from 4 to open, then 86% of the additional links will be spurious, compared with 69% if the span is changed from two to four. From this it would appear that the words closest to the node

are more likely to provide valid links. Of course, these figures hold true only for this particular concordance, as evidence from the evaluation exercise, presented in Chapter 9, will highlight. In addition, the fact that a wider span produces a larger number of spurious links does not change the fact that *valid* links are still being added as the span is increased; it is simply the proportion of spurious to valid links which is changing.

4 → open	86%
4 → 5	84%
3 → 4	77%
2 → 4	69%

Table 6.7
Effect of Span Change on Valid Link Formation

Availability of Context

It should be borne in mind that many of the 'exchange' concordance lines do not provide enough context for the larger spans and that this could account, at least to some extent, for the decrease in the number of extra bonds identified as the span is increased (Table 6.6). This mismatch can be examined in greater detail by identifying which slots are actually available in each line and then collating the results:

Slot (left)	No. of Lines	Slot (right)	No. of Lines
−8	64	+8	5
−7	122	+7	31
−6	162	+6	78
−5	175	+4	174
−4	176	+4	174
−3	176	+3	176
−2	176	+2	176
−1	176	+1	176

Table 6.8
Available Slots in 'exchange' concordance

This table tells us that there are only 64 lines which supply sufficient context for there to be a word eight slots to the left (−8) of the node word, only 122 lines which have a word seven slots to the left, and so on, up to eight slots to the right (+8), which is to be found in a mere 5 lines. The range of slots from −4 through +3 are to be found in all 176 lines, however, and in only two lines is the +4 slot missing. This adds weight to the argument, expressed above, that the majority of lines will provide a ±4 span and might also be a factor in the popularity of ±4 span collocate analysis, since early work in this field will also have been done on fixed-span concordances. As evidenced by the link word list for the open span, the number of truncated words present in the concordances is not inconsiderable. When the entire line is used, it is obvious that if partial words are present in the context then they will be candidates for linkage, but they can also be expected to fall within the context provided by a fixed-size span, if it is large enough, and can therefore create spurious links for span sizes other than open. A cursory inspection of the link word list for ±7 words confirms this expectation, as it contains fifteen obviously truncated entries. An example is 'f', which is present as the last 'word' in these lines:

```
(13)  horses, beads and cloth came south in exchange. These societies were so far f
(43)  W YORK - Prices on the New York Stock Exchange staged a blue- chip rally F
(81) f papers. "Well, there is the Rummidge exchange, but you wouldn't be intereste
```

```
(115) mena's trip has sparked a sharxpublic exchange between him and Velasco. The F
```

An additional potential problem, although no cases have so far been observed, is that the truncated form is itself a valid word which occurs on the wordlist and that a false link is therefore instantiated. While this is going to be a rare occurrence, it nevertheless represents another drawback of using a formally defined element as input to the software.

A Note on Stopwords It is worth noting here how the fixed span interacts, on a purely mechanical basis, with the stopword facility. In procedural terms, the span is applied to the concordance line first and then the stopwords are removed. This has ramifications for the type of link which can be detected, as can be seen in the lines for 'exchange' in Figure 6.3 below, which all additionally contain the item 'goods'. Using the strategy outlined above, only lines (1) and (3) will be linked via the item 'goods' (and only then if raw or relative link type is used), since 'goods' in line (2) lies outside the ±4 span delimited by the > and < signs.

```
1: conomy changed from> one based upon the exchange of goods and services< to one b
2: mpete if we are> to earn enough foreign exchange to buy the primary< goods we st
3: re businessmen of> various sorts met to exchange goods, property or money<. Afte
```

Figure 6.3
'exchange' plus 'goods' Concordance: stopwords included in span

If the alternative strategy were used, that is, the stopwords were removed first, and then the spans were extracted, the non-stopword items could be allowed to 'shuffle up'. Assuming that 'and', 'of', 'or', 'the' and 'to' are stopwords, the alternative procedure would allow 'goods' to form an absolute link between lines (1) and (3) (as 'goods:1') and a relative link for all three lines as 'goods:+', since it would then fall within the specified span delimited by the < sign. This can be seen in Figure 6.4, where all the words in bold type are now included in the span.

```
1: ... exchange of goods and services to one b<
2: ... exchange to buy the primary goods we< st
3: ... exchange goods, property or money. Afte<
```

Figure 6.4
'exchange' plus 'goods' Concordance: stopwords excluded from span

The justification for including the stopwords in determining the span contains two main threads. Firstly, as stated in the above section on Span Size, the purpose of the fixed span is to provide a contrast to the results achieved using the open span, yet the alternative methodology just described would have the effect of reducing any difference between the two span types since it would potentially allow any 'fixed' span to encompass the entire concordance line. Secondly, one of the verification methods, described in a later chapter, which has been applied to the output of *cohort* involves the identification of significant collocates within the lines selected on the basis of observed:expected frequencies of co-occurrence, an analysis which is far more complicated to perform and error-prone if the collocational context of the node is not (truly) fixed.

As an alternative to identifying the links within KWIC concordances, it would be possible, if the entire text of the corpus were accessible, to extract spans of any desired length straight from the corpus or even to use the same unit as for textual abridgement – the sentence. In order to do this it would be necessary to integrate the abridgement algorithm into a corpus retrieval system and, in the latter case, to use a corpus which had been parsed into sentences. Sadly, for the reasons presented in Sections 1.4 and 4.2.2., neither of these resources were available when this study was initiated. The benefit of using full-sentence concordances would be that the sentence is a naturally-occurring 'span' and that it would be preferable to use this instead of applying an artificial window of n words either side of the node. The task of selecting examples for use in dictionaries might also be simplified if the examples were in fact entire sentences.

Since the main motivation for using a fixed span is to compare the results with those obtained using the open span, the choice of a span of four seems the most logical, as it is known to be a reasonable compromise between comprehensive coverage and processing overheads and, if we take the 'exchange' concordance to be representative, a span of ±4 can be extracted from most concordance lines.

6.6. Conclusion

In this chapter we have dealt with each of the parameters to the concordance line selection system in turn. Each one has been defined, its possible values explored and the effects of altering it investigated. It was mentioned earlier that certain combinations of parameters might interact more usefully than others and in the following chapter this factor will be more fully examined.

Chapter 7

Interaction of Parameters

7. Interaction of Parameters

7.1. Introduction

The number of possible combinations of the parameters described in the previous chapter is not inconsiderable. Assuming that the set of possible values given earlier is used, the potential number of combinations can be determined by multiplying together the number of possible values of each variable, as follows:

Parameter	Possible Values	No. of Possible Values
Link Threshold	1, 2, 3, 4, 5, 6	6
Link Type	Raw, Absolute, Relative	3
Span Size	±4, open	2
Stopword List	bt, btb, arts-prons, top 50, top 100, top 150, zero	7

Table 7.1
Combination of Parameters

This gives ($6 \times 3 \times 2 \times 7 =$) 252 potential combinations, which does not include the possibility of using sentence-length concordances or fixed-size context of any size other than ±4. If these two further possible values of the Span Size variable are included in the calculation, then the total increases to 504. Had POS tagging been included (see Section 1.4 for reasons why it was not) as a binary (on/off) option, this would have brought the total to 1,008 different parameter combinations.

Complete scientific rigour would demand that a set of outputs should be created and assessed for each combination. Since four variables are involved, though, the results would be hard to interpret and more difficult still to present. In addition, to make this test fully rigorous it would need to be applied to several sets of concordance lines, adding yet another dimension. It therefore seems more sensible to direct one's efforts towards establishing the most generally applicable guidelines for the optimal combinations of parameters. It has already been demonstrated that the data is not just a seemingly random set of numbers obtained from a laboratory experiment and it is clear that the individual

parameters can interact in ways which favour particular combinations. It is reasonable then to assume that the predictable characteristics of the concordance lines can be exploited in order to exclude some of the many possible combinations of parameters. There would be little value, linguistically speaking, in setting a high link threshold, using a large stopword list, a fixed-size span and absolute positioning, since intuition would inform us that few, if any, concordance lines would be selected, although this configuration of parameters would be a means of identifying lines which become bonded solely because they are near or perfect duplicates. The fact that some combinations are unlikely to prove useful is attributable to the existence of collocational 'profiles'† for each word examined. Were it not for this feature of the language, the approach described herein would be of no linguistic or practical interest.

Collocational profiles represent one of the predictable features of the language and make the analytical methodology described herein workable, but in contrast to them there is a degree of randomness in the input (the concordance) in the form of those items, lexical and grammatical, which do not form part of the collocational profile of the node word in question. These constitute a 'wildcard' factor in the input, since there is no way to predict their presence. Such items, because they are not expected to collocate with the node, are unlikely to contribute significantly to the link score of the particular line in which they appear, since, by definition, collocates occur repeatedly and significantly with their node word. Of course, such wildcard items only need to occur twice in order to form a link, but this level of occurrence would be far too low for these items to be classed as collocates. The effect of such words on the bond score will depend on other factors: they may be removed entirely if the link type (positional specification) is strict enough, or cancelled out by the action of the link threshold, if this is raised above one, or they may even be caught by the stopword list.

---

† The concept of collocational profiles was introduced in the discussion of the AVIATOR project in the earlier chapter describing the development of the concordance line selection software, *cohort*. See Renouf 1994, Collier & Pacey 1996 for further discussion of collocational profiles and their practical applications.

Through an awareness of certain characteristics of the input, it becomes possible to predict, despite the partially random nature of the input, particular combinations of parameters which might be used to detect other linguistic phenomena. These characteristics can be concrete, such as the fixed size of the concordances, or more abstract, such as collocational patterning. One such example of this has already been introduced in the section on the Link Threshold in the previous chapter, where it was shown that setting a high link threshold could highlight lines containing idioms or other fixed phrases.

7.2. Effect of Parameters

In calculating a score for individual concordance lines, two components of the *cohort* system are centrally involved prior to the final calculation of the number of bonds: the wordlist and the matrix. As was noted at the beginning of the previous chapter, the various parameters come into play at different stages and thus have an effect on either the wordlist, the matrix or the final bond score. Let us now look briefly at the influence exerted by each of the parameters on these components:

	Affects Wordlist	Affects Matrix	Affects Bonds
Link Threshold	No	No	Yes
Positional Specification	Yes	via wordlist	via matrix
Stopword List	Yes	via wordlist	via matrix
Span	Yes	via wordlist	via matrix

Figure 7.1: Range of Effect of Parameters

From this it can be seen that all but one of the parameters, the Link Threshold, have an influence on the creation of the wordlist. Since the Link Threshold acts upon the matrix to produce the list of bonded lines, it follows that once a given wordlist has been created there is only one possible matrix which can be built from it, that is, none of the parameters can affect the transformation of the wordlist into the matrix.

In the introduction to this chapter, the possibility was mentioned of running the software using every possible permutation of parameters and this has indeed been done. For the reasons stated earlier, however, no detailed analysis of each single result will be presented here, but there is just sufficient room to present the now-familiar metrics of links and bonds for each combination. In addition, a standard deviation has been calculated on the bond scores obtained using each set of parameters. This is intended to convey an idea of the variation in bond score values across the lines, which ties in with the point made in the Parameters chapter about the desirability of obtaining a wide range of scores. The metrics were obtained by iteratively running the program over the 'exchange' concordance, incrementing, or making stricter, each parameter in turn in a series of nested loops, which might be represented symbolically as:

FOR each value of Link Threshold

FOR each value of Positional Specification

FOR each value of Span Size

FOR each value of Stopword List

RUN with these values

The data were then sorted on the Link Type, Span Size, Stopword List and Link Threshold (in that order of priority) to produce a table 252 lines in length which, for the sake of brevity here, can be found in Appendix 3.

7.3. Focusing on Links

7.3.1. Ignoring the Link Threshold

What emerges from the table in Appendix 3 is that the number of link words and total links never changes for a given combination of Link Type, Span Size and Stopword List. As an example, examine the link information for all the cases of 'Abs 4 artsprons', i.e. the first six lines of the table. In all instances, 135 link words and 437 total links are

identified, amplifying the point made earlier that the link threshold is applied to the completed matrix and does not therefore have any effect on the formation of links. Since this consistency holds for all the other combinations of the three remaining parameters, we can infer that there are actually far fewer than 252 possible wordlist and matrix configurations. The real number is arrived at by ignoring the influence of the six possible values of Link Threshold and is therefore ($3 \times 2 \times 7 =$) 42† (or $252 \div 6$). This insight enables us to strip away a layer from this rather unwieldy table and, for the time being, to ignore the role of the link threshold. By removing the columns relating to the threshold and the bonds, a far more tractable table, 7.2 below, is created, containing only 42 items and based on the three-way relationship between Stopwords, Span and Link Type. Since these are the parameters which are active during the creation of the wordlist, they will be henceforth referred to as the wordlist parameters.

Table 7.2: Combination of Parameters - Link Threshold removed

Item No.	Link Type	Span Size	Stopword List	Number of Link Words	Total Links
1	Raw	open	zero	236	1361
2	Rel	open	zero	251	1306
3	Raw	open	arts-prons	221	1031
4	Abs	open	zero	265	944
5	Rel	open	arts-prons	225	936
6	Raw	4	zero	154	883
7	Rel	4	zero	166	843
8	Abs	4	zero	177	690
9	Raw	4	arts-prons	140	638
10	Abs	open	arts-prons	197	598
11	Rel	4	arts-prons	145	584
12	Raw	open	btb	159	571
13	Raw	open	top50	188	561
14	Rel	open	btb	140	488
15	Rel	open	top50	166	448
16	Abs	4	arts-prons	135	437
17	Raw	open	bt	145	422
18	Raw	open	top100	147	410
19	Raw	open	top150	128	350
20	Raw	4	btb	92	345

† As readers of The Hitchhiker's Guide to the Galaxy will be interested to hear.

Table 7.2: Combination of Parameters - Link Threshold removed					
Item No.	Link Type	Span Size	Stopword List	Number of Link Words	Total Links
21	Rel	open	bt	125	339
22	Rel	open	top100	123	328
23	Abs	open	btb	107	317
24	Raw	4	top50	108	308
25	Rel	4	btb	81	302
26	Rel	open	top150	103	279
27	Rel	4	top50	93	252
28	Abs	4	btb	73	238
29	Raw	4	top100	84	236
30	Raw	4	bt	83	233
31	Abs	open	top50	93	227
32	Raw	4	top150	72	200
33	Rel	4	top100	71	195
34	Abs	open	bt	76	193
35	Rel	4	bt	69	189
36	Abs	open	top100	73	186
37	Abs	4	top50	67	172
38	Rel	4	top150	60	166
39	Abs	open	top150	64	164
40	Abs	4	top100	51	139
41	Abs	4	bt	51	138
42	Abs	4	top150	43	119

The lines of Table 7.2 have been sorted in descending order of total links in order to make any correlation which exists between total links and link words more apparent. On the whole, it is indeed observable that the lower the number of link words, the fewer total links will be generated and this is confirmed by the Pearson's Correlation Coefficient of 0.92, which is indicative of a highly† significant correlation between link words and total links. The closeness of the correlation can be seen graphically in the next figure:

† Based on the approximation of the transformed correlation coefficient to Normal distribution as described in the next chapter. For n=42 and r=0.92, t is equal to 10.13, which is significant at the 1% level.

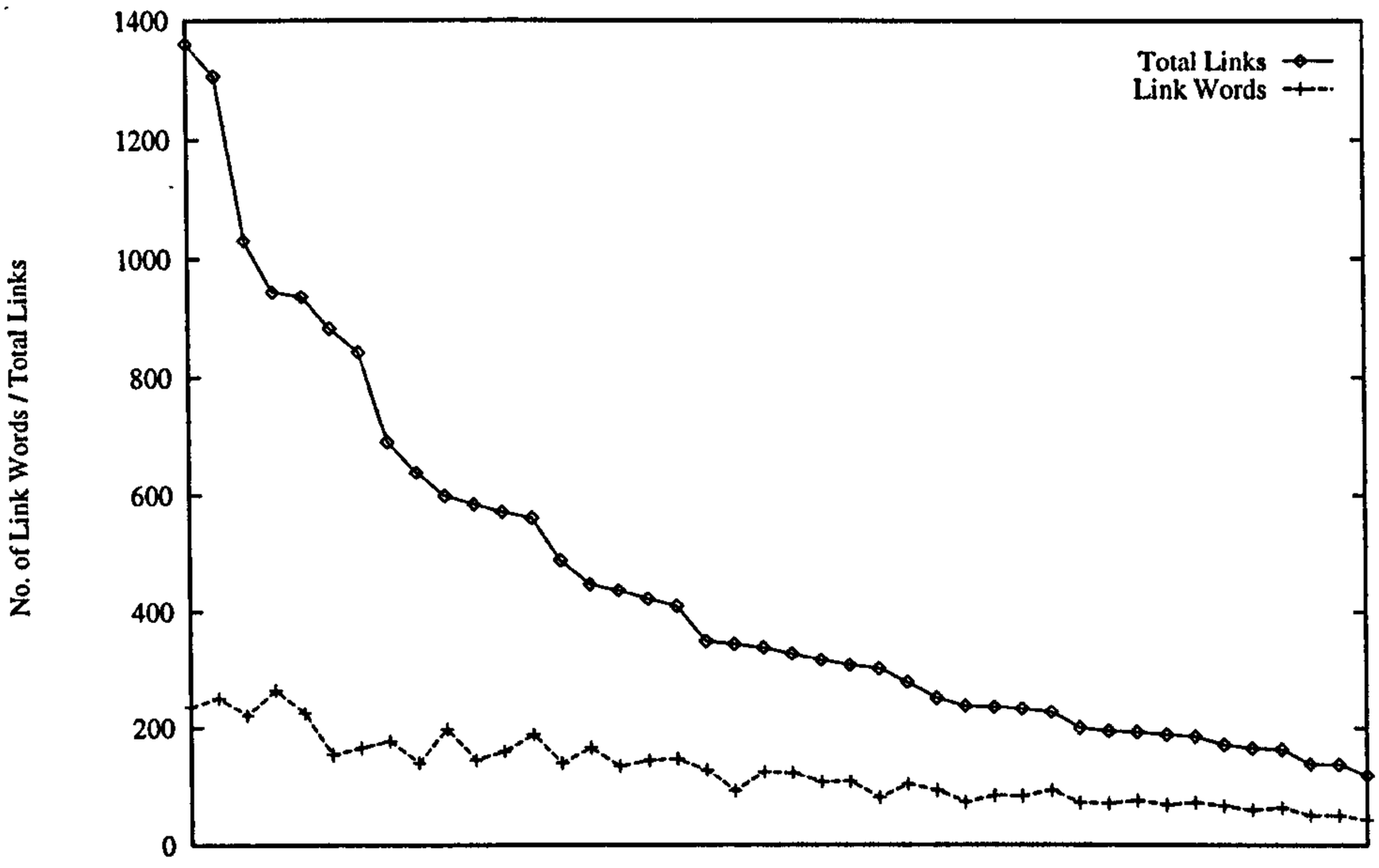

Figure 7.2
Total Links vs Link Words for each Wordlist Parameter Combination

As might be expected, the greatest number of total links are formed when the least strict parameters, **Raw**, **Open** and **Zero**, are used. Item 2 on the list, of course, produces fewer total links, but the number of link words is higher than for item 1. This can be attributed to the use of the stricter positional specification, **Rel**, for item 2, since either of the link types other than raw will tend to increase the number of types in the wordlist, because of the attachment of the positional information to each token in the concordance. Because the occurrences of the word are being diluted by the addition of the positional marker, however, their chances of forming links are reduced. In effect, several types are created from the one *Ur*type, differing only in their positional specification. As an example, here is the wordlist entry for the word 'information', extracted from the 'exchange' concordance using raw positioning:

```
information 25 56 87 152 155
```

Notice that it establishes a link between five lines. If this is now compared with the entries for the same word from a wordlist which has been built using the absolute link

type:

```
information:1 87 155
information:2 25 152
```

it can be seen that the stricter positional specification has caused one of the links (56) to be lost because 'information' did not occur in either the +1 or +2 slot in line 56. This behaviour is confirmed by item 4 of Table 7.2, which reveals that the absolute link type has increased the number of link words relative to the two other items (1 & 2) which use the same span and stopword list, but that the number of total links has gone down.

7.3.2. Stopwords and Link Type

The observation that the use of a stricter positional specification increases the number of link words holds for all combinations of wordlist parameters which use the 'zero' stopword list (items 1, 2, 4, 6, 7 & 8 in Table 7.2). If a different stopword list is substituted, however, this relationship no longer holds. Using the 'arts-prons' list, for instance, results in the most link words being identified when the relative link type is employed, with the absolute positional specification resulting in the fewest link words. If a larger stopword list is used, then the situation changes again. The 'top50' list, and all subsequent stopword lists in fact, cause raw links to form the greatest number of link words, followed by relative links and then absolute links. This is summarised in Table 7.3 below which presents the stopword lists in ascending order of size:

Stopword List	No. of Link Words greatest → smallest
zero	Abs, Rel, Raw
arts-prons	Rel, Raw, Abs
top50	Raw, Rel, Abs
top100	Raw, Rel, Abs
top150	Raw, Rel, Abs
btb	Raw, Rel, Abs
bt	Raw, Rel, Abs

Table 7.3
Effect of Stopwords and Positional Specification on Link Formation

Naturally, this raises the question of why this should be so. The answer would seem to lie in the relationship between the likelihood of a word changing its position relative to the node word and whether it is a stopword. The next table summarises, for each type of link, the number of link words which remain in the wordlist as each stopword list in turn is applied. In addition, it supplies a percentage figure, calculated by dividing the number of remaining words by the original total as represented by the figures for the 'zero' stopword list. Thus for 'arts-prons' and raw links, 140 link words remain, which is 91% of the 154 words allowed by the 'zero' list.

Stopword List	Span	Raw Links		Relative Links		Absolute Links	
		n	%	n	%	n	%
zero	4	154	100	166	100	177	100
arts-prons	4	140	91	145	87	135	76
top50	4	108	70	93	56	67	38
top100	4	84	55	71	43	51	29
top150	4	72	47	60	36	43	24
btb	4	92	60	81	49	73	41
bt	4	83	54	69	42	51	29
zero	open	236	100	251	100	265	100
arts-prons	open	221	94	225	90	197	74
top50	open	188	80	166	66	93	35
top100	open	147	62	123	49	73	28
top150	open	128	54	103	41	64	24
btb	open	159	67	140	56	107	40
bt	open	145	61	125	50	76	29
Average		–	71	–	62	–	48
% fall-off		–	0	–	13	–	32

Table 7.4
Proportion of each Link Type retained using different Stopword Lists

In the numerical (n) columns, the different orderings presented in Table 7.2 are reiterated. The percentage columns, by comparison, indicate that the *proportion* of link words retained is consistently lower for positionally-restricted link types than for raw links and that absolute links are affected to a greater extent than relative ones. The increased downward pressure on link words as the link type becomes stricter is summarised in the 'Average' row, which shows a fall-off 13% and 32% from the raw link average of 71% to the

relative and absolute averages, 62% and 48% respectively.

It is interesting to note that the span parameter appears to have little influence on the proportion of the link words which are retained, although, as might be expected, the number of link words is generally lower for the smaller span. The interaction between the span and the link type parameter is worthy of investigation, however, and this will be undertaken in the next section.

7.3.3. Span and Link Type

Since the relationship between the number of link words and the number of total links has already been explored in a previous section and the role of the stopword and span type parameters in that relationship has been identified, this and subsequent sections will concentrate on the changes in the total number of links which are brought about by various wordlist parameter combinations. Of course, if any important effects on the link words are uncovered these will still be treated separately.

If the information about the total links obtained using the different wordlist parameter combinations is rearranged so that the number of links for the two different spans are adjacent, a more detailed picture of the contribution of the span parameter can be obtained, focusing in particular on its interaction with the link type. In Table 7.5 below, this comparison has been made by setting side by side the total number of links allowed by each span value for the various combinations of link type and stopword list. The value for the ±4 span has then been divided by the figure for the open span to arrive at a percentage of the links which are retained, such that a low figure will indicate that a large proportion of links are lost when the span size is reduced. As shown by the first line of the table, for instance, using the absolute link type and **top50** stopword list results in 227 links when the open span is used and 172 when the fixed span is used, giving a ratio of 75.8% (100 × 172 ÷ 227) of links retained. The table entries are then arranged in descending order of the percentage score.

Other Parameters	No. of Total Links Open	4	% Retained
Abs top50	227	172	75.8
Abs btb	317	238	75.1
Abs top100	186	139	74.7
Abs arts-prons	598	437	73.1
Abs zero	944	690	73.1
Abs top150	164	119	72.6
Abs bt	193	138	71.5
Raw zero	1361	883	64.9
Rel zero	1306	843	64.5
Rel arts-prons	936	584	62.4
Raw arts-prons	1031	638	61.9
Rel btb	488	302	61.9
Raw btb	571	345	60.4
Rel top100	328	195	59.5
Rel top150	279	166	59.5
Raw top100	410	236	57.6
Raw top150	350	200	57.1
Rel top50	448	252	56.2
Rel bt	339	189	55.8
Raw bt	422	233	55.2
Raw top50	561	308	54.9

Table 7.5
Effect of Varying Span and Link Type on Total Links

The most obvious feature of the table is that the seven parameter combinations involving the absolute positional restriction have the highest ratio scores, occupying all the first seven slots with an average score of 73.7%, compared with 64.2% for the overall average, 60% for relative links and 58.9% for raw link types. As suggested by the closeness of their average values, these other link types are interspersed fairly evenly throughout the remainder of the table. This would seem to indicate that the number of links is affected less by the span parameter when absolute links are being used.

Not unsurprisingly, the parameter combinations involving the smaller stopword lists also figure more strongly near the top of the table, but this will be covered in greater detail in the next section.

It was demonstrated in the previous chapter that the ±4 span accounted for approximately 65% of the total links found in the open span when the other parameters were set at their

least strict values, raw links and no stopwords. Given that a stricter link type is being used here, it seems likely that more of the links will have occurred within the ±4 context of the node word, because the stricter setting has made it more difficult for them to be repeated unless they are part of the collocational profile of the node word and are therefore occurring repeatedly in the same slot. Put another way, the links are unlikely to be present at the same absolute position unless they are collocates and if they are collocates they are likely to have occurred within four words of the node word. This hypothesis is borne out by the figures in the above table for the parameter combination 'Abs zero', which indicate that 73% out of the 944 total links are to be found within the four-word span, a marked increase on the 65% value for unrestricted link position. This can be seen diagrammatically in the following graph, which plots each position of the span against the number of links which were established in that position.

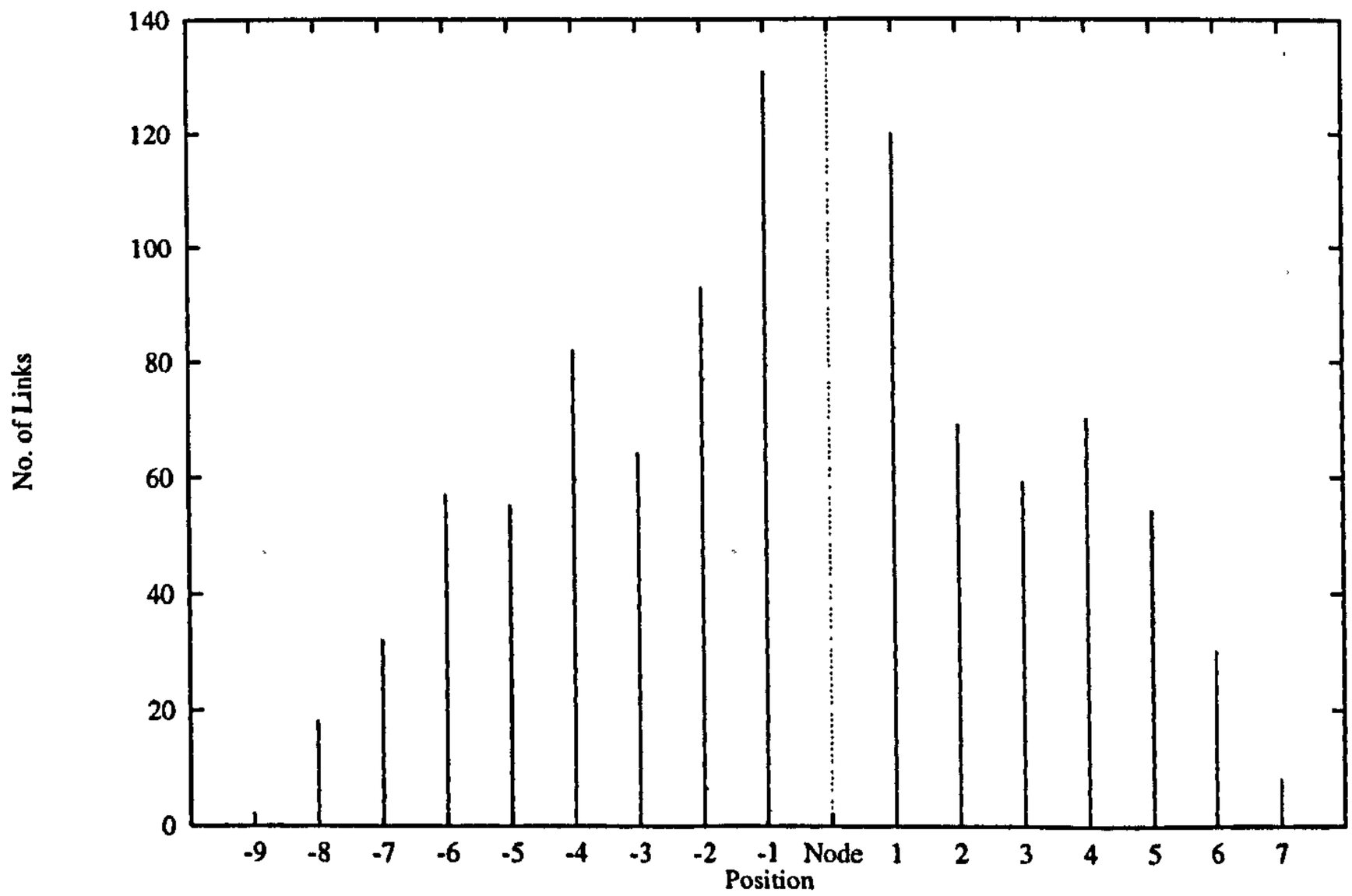

Figure 7.3
Link Position vs Frequency

This plot shows the expected fall-off in the number of links in the slots further away from the node word, but also reiterates the point made earlier that many links are occurring

outside the ±4 span.

7.3.4. Stopwords and Span

The final side of the triangle of parameters which act upon the wordlist is completed by the relationship between the Stopword List and the Span. In order to isolate the effect of this combination of variables on the number of links, the figures for the individual link types from Table 7.5 were averaged together to produce Table 7.6 which is presented in ascending order of the number of links.

Stopword List	Total Links Span		% Links Retained
	Open	4	
top150	264	162	61.4
top100	308	190	61.7
bt	318	187	58.8
top50	412	244	59.2
btb	459	295	64.3
arts-prons	855	553	64.7
zero	1204	805	66.9
Average			62.4

Table 7.6
Effect of Stopwords and Span on Total Links – Average of all link types

Other than the obvious difference in the number of links established for each list there is little to highlight any particular entry in the table and on the whole it follows the trend, identified in the 'Stopwords' section of the previous chapter, that a smaller stopword list allows more links, with the caveat regarding the **bt** and **btb** lists. As the total number of links grows, there is only a slight increase in the percentage of links retained when the span is reduced.

Comparing the results for the different spans, the figures echo the earlier conclusion regarding the contribution of the span parameter, with the average percentage retained being 62.4.

7.3.5. Summary of Effects

The comparative effect on the link words and total links of the wordlist parameter combinations can be seen in the graph in Figure 7.4. It differs from Figure 7.1, which was in descending order of total links, in that it presents the link data in related groups of parameter combinations and is therefore labelled along the horizontal axis with the appropriate link type and span parameters, labelling of stopwords being omitted for clarity. Thus the seven points at and to the right of the *Raw/4* label all represent results obtained using the raw link type with a ±4 span; each point then corresponds to one of the seven stopword lists, which are presented in ascending order of size, namely **zero**, **arts-prons**, **top50**, **top100**, **top150**, **btb** and **bt**. After these come the seven points for raw links and open span starting at the label *Raw/O* and so on.

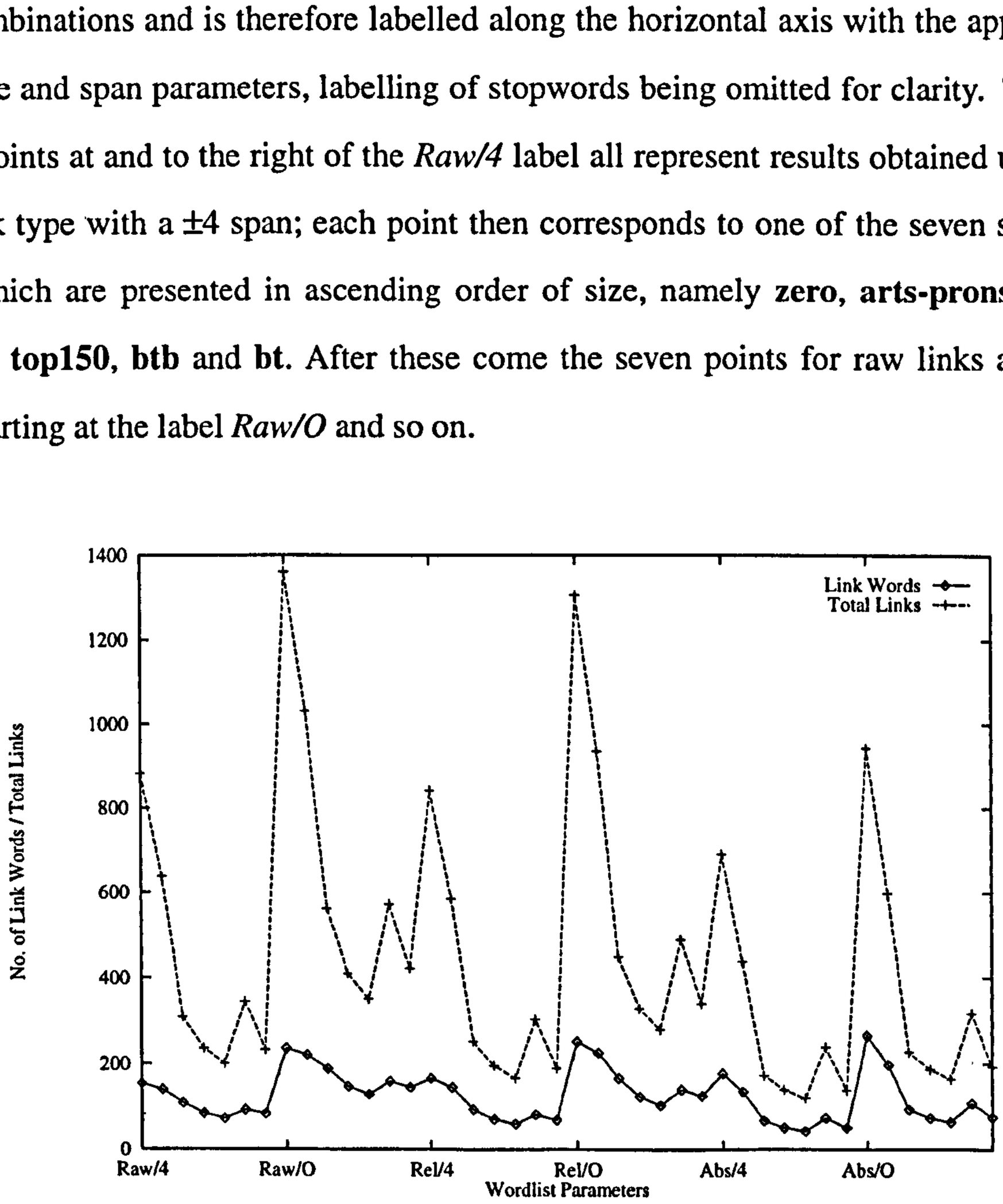

Figure 7.4
Effect of Wordlist Parameters on Links

The regular effect of the different stopword lists is clearly visible, especially in the upper Total Links plot, as six repeated patterns of two peaks, one large, one small, caused by the decrease in links as the frequency-based stopword lists increase in size with a final blip

caused by the **btb** and **bt** lists which are restrict link formation less, relative to their size. In addition, the contrast between the two different spans can be easily seen by comparing pair-wise the six patterns mentioned above, i.e. Raw/4 against Raw/O etc.

As one would expect, overall the graph shows the close correlation between the number of link words and the total number of links first seen in Figure 7.1. In this arrangement of the data it is possibly even clearer, with the troughs and peaks coinciding exactly, although the amplitude of the Total Links plot is considerably greater, especially where the open span is used.

7.4. Focusing on Bonds

So far, this chapter has concentrated on the interaction of those parameters which are involved in the creation of the wordlist. It has been demonstrated that for a given wordlist there is only one possible matrix, since the process of creating the matrix from the wordlist is independent of all the variables discussed here. By examining the point in the process at which each parameter comes into play, it was possible to remove the Link Threshold from consideration temporarily and thereby discuss a less complicated set of combinations of parameters in terms of their effect on the number of link words and total links created. In the remainder of this chapter, we shall reintroduce the Link Threshold and look at its effect upon the process of transforming links into bonds.

To recap on the role of the link threshold, it is applied to the contents of the matrix in order to determine whether a particular pair of lines share sufficient links for a bond to be established between them. If this condition is met, then the bond score for each of the lines, stored on a list which is external to the matrix, is incremented by one. Once the whole matrix has been scanned for potential bonds, the scores from the bond list are attached to the original concordance lines which are then ranked according to the number of bonds acquired by each line.

Since a link threshold as low as one is allowable in the concordance line selection system, intuition suggests that a large number of total links will result in many bonds being established, since if the link threshold is one, then each link creates a bond. The correlation between total links and bonded lines should therefore be a strong one. It is unlikely ever to be a perfect correspondence, since there will always be bonds made up of more than one link, otherwise there would be a direct relationship between the number of links present in the wordlist and the number of non-zero matrix cells. In Table 7.7 below, the Total Links figures for the 42 wordlist parameter combinations are shown in ascending order. Alongside these are the figures relating to the number of lines which are bonded when link thresholds of between one and six links are applied to the 'exchange' concordance.

Total Links	Link Threshold 1	2	3	4	5	6	Wordlist Parameters
119	88	18	2	2	0	0	Abs 4 top150
138	98	16	4	2	0	0	Abs 4 bt
139	99	18	6	2	0	0	Abs 4 top100
164	102	18	9	2	0	0	Abs open top150
166	115	21	2	2	0	0	Rel 4 top150
172	112	20	8	4	0	0	Abs 4 top50
186	111	18	11	4	0	0	Abs open top100
189	122	19	4	2	0	0	Rel 4 bt
193	112	16	11	4	0	0	Abs open bt
195	126	21	6	2	0	0	Rel 4 top100
200	122	27	2	2	0	0	Raw 4 top150
227	122	20	11	6	2	2	Abs open top50
233	132	25	4	2	0	0	Raw 4 bt
236	134	27	6	2	0	0	Raw 4 top100
238	134	44	10	2	0	0	Abs 4 btb
252	142	27	8	4	0	0	Rel 4 top50
279	144	28	11	2	0	0	Rel open top150
302	154	51	12	2	0	0	Rel 4 btb
308	151	33	8	4	0	0	Raw 4 top50
317	145	45	15	8	0	0	Abs open btb
328	154	30	13	4	0	0	Rel open top100
339	152	29	13	4	0	0	Rel open bt
345	159	61	10	2	0	0	Raw 4 btb
350	153	35	13	2	0	0	Raw open top150

Total Links	Link Threshold 1	2	3	4	5	6	Wordlist Parameters
410	162	38	15	4	0	0	Raw open top100
422	158	40	15	4	0	0	Raw open bt
437	173	74	21	4	4	4	Abs 4 arts-prons
448	164	37	13	8	2	2	Rel open top50
488	168	77	28	10	0	0	Rel open btb
561	172	53	15	8	2	2	Raw open top50
571	172	95	34	8	0	0	Raw open btb
584	175	107	35	4	4	4	Rel 4 arts-prons
598	174	83	27	11	4	4	Abs open arts-prons
638	176	132	37	4	4	2	Raw 4 arts-prons
690	175	127	59	16	6	4	Abs 4 zero
843	176	162	90	29	8	4	Rel 4 zero
883	176	173	108	31	8	2	Raw 4 zero
936	176	152	65	20	4	4	Rel open arts-prons
944	176	140	63	19	11	6	Abs open zero
1031	176	166	92	19	4	2	Raw open arts-prons
1306	176	176	142	68	26	8	Rel open zero
1361	176	176	158	108	22	6	Raw open zero

Table 7.7
Effect of Link Threshold on Bond Formation

There appears to be a high degree of correlation between several of the Link Threshold columns and the Total Links column and this is confirmed by the Pearson's Correlation Coefficient scores shown in Table 7.8, which are all highly significant, although the worth of the lines selected when the link threshold is six is questionable, since the degree of bonding is so low and there are so many combinations which result in no lines being bonded.

Link Threshold	r
1	0.778832
2	0.950493
3	0.9468
4	0.818683
5	0.860199
6	0.84777

Table 7.8
Pearson's Correlation Coefficient (r) for Total Links vs Bonded Lines

7.5. Conclusions

In this chapter the important distinction between those parameters which affect the contents of the wordlist, and thereby the matrix, and the remaining parameter which controls bond formation has been established. It has been shown that there is a strong correlation between the number of link words found in the wordlist and the total number of links formed. In addition to this, the relationship between links and bonds has been explored in depth.

On the basis of the above findings, what, then, can be said of the optimal settings for the four parameters described here? If the intention is to identify the most generally applicable settings, then it is necessary to select values which will bring about the highest degree of bonding, since this will be most likely to provide sufficient information to rank the concordance lines. If the parameters are so strict that the majority of lines are not bonded, then the selection will be coarse-grained; basically a line will be selected or not, according to whether it is bonded. Ideally there should be a wide range of bond scores assigned, possibly with zero bonds forming part of that range, so that a fine-grained distinction can be made and the lines can ultimately be selected by the human user, who can decide how far down the ranked concordance to proceed. If the range of bond scores is limited and many lines score no bonds, on the other hand, then the final selection is essentially made by the computer on the basis of whether a line is bonded or not. The fine-grainedness of the bond scores is not, then, in itself a reflection of the typicality/ centrality/ representativeness of the lines which are forming bonds, but is rather a means to the end of providing the corpus researcher with a tool which is as finely calibrated as possible.

To achieve the largest number of bonded lines and the greatest variation in bond scores, the results presented so far suggest that a fairly liberal link threshold should be used, combined with a small stopword list. This can be confirmed by calculating standard deviations for the scores assigned by each set of parameters and in the table which follows, the ten parameter combinations with the highest standard deviations are listed. The mean

number of bonds per line (total bonds over number of bonded lines) is also given.

Parameters	Total Bonds	Mean Bonds	Bonded Lines	SD
zero 1 fixed raw	15,556	88.3864	176	34.3318
zero 1 open rel	16,462	93.5341	176	31.2
zero 1 open raw	21,602	122.739	176	31.162
arts-prons 1 open raw	12,404	70.4773	176	29.5764
zero 2 open raw	9,208	52.3182	176	29.3662
zero 1 fixed rel	10,732	60.9773	176	28.3098
arts-prons 1 open rel	8,164	46.3864	176	23.119
btb 1 open raw	4,618	26.2386	172	21.9576
arts-prons 1 fixed raw	7,100	40.3409	176	21.9425
zero 2 open rel	5,282	30.0114	176	20.3207

Table 7.9
Top ten Parameter Combinations by Standard Deviation

The strength of the standard deviation scores (SD)† indicates that there is a wide range of bond scores present in the output for the parameters shown. In addition, nearly all the sets of parameters shown here have caused all 176 lines to be selected. These conditions hold for many more of the parameter sets, even those with link thresholds higher than two. It should therefore be possible to retain the necessary granularity, while at the same time focussing on features which call for higher link thresholds such as compounds, phrasal verbs and other fixed strings.

This concludes the examination of the effect of the different parameter combinations. In a subsequent chapter, the various automatic analyses will be compared and contrasted with the manual analysis carried out by a group of experienced corpus users, but first we present some examples of the output from *cohort*.

The SD score gives a measure of the variation across the bond scores in a better way than the mean bond score. If A is the set of scores (1 2 3 4 5 6 7 8 9 10) and B the set of scores (5 5 5 5 5 5 5 5 5 5), then mean(A) is 5.5 and mean(B) is 5, i.e. they are close in value. SD(A), however is around 3, while SD(B) is 0 because there is no variation in the values of B.

These are population standard deviation scores calculated using N.

Chapter 8

Output from the Software

8. Output from the Software

8.1. Introduction

In Chapter 7, the interactions of the various parameters to the *cohort* system were explored, providing some useful insights into the combinations of parameters which might prove most useful in generating a fine-grained ranking of the elements of the concordance. These parameter combinations resulted in the majority of the lines being selected (i.e. achieving bonds with other lines) and also allowed a good range of bond scores, as measured by the Standard Deviation across all the bond scores in the concordance. In the next chapter, a full evaluation of the output from *cohort* is carried out, based on comparisons with a manual analysis of concordance data. For now, however, we will examine the output from the parameter combinations listed in Table 7.9. This will also serve to reacquaint the reader with the output from the software, as it has been some time since this was first presented.

8.2. The Output

At the end of the previous chapter, ten sets of parameters were identified which fulfilled the criteria of having many bonded lines and good variation in bond scores. In general, these combinations employed liberal link thresholds, raw or relative link types and small stopword lists, these being factors which were isolated in the course of Chapters 6 and 7 as being likely to generate output with the desired characteristics. The remainder of this chapter will present the output for each of the ten settings.

The format of the output from *cohort* is very similar to the conventional concordances which formed the input to the process. Each line in the output is prefixed with two figures, the first of which is simply the line number in the original concordance. It is zero-padded to four digits in order to help preserve the formatting of the concordance lines, so that line 99 is labeled as 0099. The second figure shows the number of bonds which the

line acquired. It is also zero-padded (to three digits) and it is this value which is used in deciding the order of presentation of the lines, in that they are ranked in descending order of the number of bonds.

For each combination of parameters, the 25 lines that scored the most bonds are presented, along with the bond and standard deviation (SD) figures given in Table 7.9. A short commentary on the results is also given. This is not intended to be a thorough-going evaluation, but rather an exploration of the features which are highlighted as interesting by the software.

8.2.1. zero/1/fixed/raw

15,556 total bonds, 176 bonded lines, SD: 34.3

```
0007 145  endeavoured to heal the wounds. In an exchange of letters with Mansholt he de
0011 144  here was unbearable. And he wanted to exchange the unbearable for the very ba
0152 139 rovides a national focal point for the exchange of information, ideas and expe
0037 139 Flanders. The printer was entranced to exchange a few of the place-names which
0060 138 ate a small plot on the worst land, in exchange for agricultural and even dome
0054 138 and manufactures to the third world in exchange for raw materials and food, is
0107 137 man gives food, care and protection in exchange for the different services the
0099 136 inder of the session was devoted to an exchange on the compatibility of religi
0161 135 the night officer and the sister would exchange a few words with us. In my fir
0092 134 he flat and go to a hotel. The rate of exchange in Denmark is heavily against
0028 132  when its shares are introduced to the Exchange, probably in January. He also
0006 131  could utter silence. In practice, the exchange of letters often takes a full
0143 129 pt to minimize the risk of a strategic exchange with the Soviet Unio~ with th
0123 129 ne was out of order; on the second the exchange was closed for a religious hol
0102 129 ld be impossible for the "fighters" to exchange roles so freely. They would be
0005 129  came from Berlin and abroad, eager to exchange the new ideas that were racing
0140 128 papers presented to the Securities and Exchange Com- mission, the multinationa
0130 128 oco had reported to the Securities and Exchange Commission. Recently, Texaco l
0097 128 id enjoyment. The justification is the exchange of ideas, and the value of thi
0025 127  to freer, cheaper and more widespread exchange of information between the ric
0150 127 rld have a chance to meet together and exchange ideas. The Vegetarian Federal
0018 127  of value, the proposition that things exchange in accordance with the amount
0064 127 conomy changed from one based upon the exchange of goods and services to one b
0002 125  Ruder, chairman of the Securities and Exchange Commission, said Britain, the
0114 124 means of production, distribution, and exchange". The prose style of the notor
```

This concordance illustrates a number of features of the node word that correspond well with linguistic intuition. Firstly, several strong collocates of 'exchange' are exemplified here: 'letters' (two occurrences), 'information' (twice), 'words' (one occurrence), 'rate' (once), 'ideas' (four times), 'goods' (once). There are also a number of phrases containing 'exchange' as an element, both lexical: 'rate of exchange', 'Securities and Exchange Commission' and grammatical: 'an exchange of X', 'in exchange for'.

The degree of bonding exhibited in the lines is reasonably high, which is to be expected given the low level of strictness imposed by the parameters used to generate this output. There are, nevertheless, some lines which are not entirely satisfactory. Line 11 (0011), for example, seems to contain very little that could be creating bonds with other lines; perhaps its high bond score is attributable to the presence of the presence of the 'to exchange ... for' construction, bolstered possibly by the use of 'wanted to...'; in terms of the lexical items in the context it is remarkably atypical – 'unbearable' is by no stretch of the imagination a common collocate of 'exchange'. A similar criticism might be levelled at line 37, which contains the unlikely collocates 'printer' and 'entranced'. The lines which contain the capitalised form 'Exchange' are also worthy of comment. Although several such lines exist, three of which contain strong phrasal constructions, the degree of bonding is far too high for this to be caused by the lexical items alone and so, since no stopwords are employed in creating this ranking, it has to be concluded that the grammatical items are largely responsible for the high bond scores for these lines.

8.2.2. zero/1/open/rel

16,462 total bonds, 176 bonded lines, SD: 31.2

```
0099 148 inder of the session was devoted to an exchange on the compatibility of religi
0092 146 he flat and go to a hotel. The rate of exchange in Denmark is heavily against
0007 146  endeavoured to heal the wounds. In an exchange of letters with Mansholt he de
0054 143 and manufactures to the third world in exchange for raw materials and food, is
0143 142 pt to minimize the risk of a strategic exchange with the Soviet Union with th
0084 142 for the walls of their private bars in exchange for a few pints of beer. Or, i
0037 138 Flanders. The printer was entranced to exchange a few of the place-names which
```

Collier

```
0140 136 papers presented to the Securities and Exchange Com- mission, the multinationa
0066 136   d an expert in Round Tableip and the Exchange of Unpleasantries. Last certai
0055 136 and other statues from the first Royal Exchange). The Library is a first-class
0167 135 ush to sell. This hit the floor of the Exchange with torrential force. The mac
0068 135 d, "is not the way to begin a cultural exchange." The incident caused the trai
0142 134 pproached the defense table, hoping to exchange a few words with them. The gu
0163 132 tinued to a imported, and there was an exchange of "light" North Sea oil for "
0044 132 a and are the same kind of judgemental exchange these ladies, as children, ove
0173 131 y another name. In the absence of real exchange controls, however, the tax aut
0040 131 N A SINGLE IMAGINATIVE GESTURE. AT THE EXCHANGE WE GET THE OLD SPECTACLE OF A
0060 130 ate a small plot on the worst land, in exchange for agricultural and even dome
0002 130  Ruder, chairman of the Securities and Exchange Commission, said Britain, the
0087 129 gned to work with the local service to exchange information, to train the loca
0053 129 and maintain the correct rate of fluid exchange within the various fluid compa
0152 128 rovides a national focal point for the exchange of information, ideas and expe
0097 128 id enjoyment. The justification is the exchange of ideas, and the value of thi
0038 127 Lou Darrow Carrington runs the foreign exchange desk for the bank's corporate
0101 126 kbrokers did by the pillars of the old Exchange. Through the west gate of this
```

This sample contains more uses of the 'conversation' meaning of 'exchange' (lines 99 & 44) than were seen in the previous set and also appears to represent the financial sense more strongly. The bond scores are just as high as in the previous sample, indeed the total number of bonds is higher. In general, there is good representation of strong collocates and phrases, although the value of line 66 seems questionable, apart from the fact that it exemplifies the 'the exchange of X' paradigm.

8.2.3. zero/1/open/raw

21,602 total bonds, 176 bonded lines, SD: 31.2

```
0092 169 he flat and go to a hotel. The rate of exchange in Denmark is heavily against
0007 164  endeavoured to heal the wounds. In an exchange of letters with Mansholt he de
0161 163 the night officer and the sister would exchange a few words with us. In my fir
0152 160 rovides a national focal point for the exchange of information, ideas and expe
0084 160 for the walls of their private bars in exchange for a few pints of beer. Or, i
0060 160 ate a small plot on the worst land, in exchange for agricultural and even dome
0054 160 and manufactures to the third world in exchange for raw materials and food, is
0011 159  here was unbearable. And he wanted to exchange the unbearable for the very ba
0163 157 tinued to a imported, and there was an exchange of "light" North Sea oil for "
0066 157   d an expert in Round Tableip and the Exchange of Unpleasantries. Last certai
0086 156 ght to harvest any farmer's fields. In exchange, they get one ninth of the cro
0146 155 re of the enormous destruction such an exchange would cause, and this awarenes
0006 155  could utter silence. In practice, the exchange of letters often takes a full
```

0049 155 al Festival, which is held in the Corn Exchange each May. Miss Gray and I had
0044 155 a and are the same kind of judgemental exchange these ladies, as children, ove
0023 154 the best way in an emergency. The SIS exchange called Boyd Stuart's home and
0036 154 Copies were burned on the London Stock Exchange and destroyed at exchanges in
0172 153 wine later, and frolicketwith an oral exchange of that, laughing over the hyd
0143 153 pt to minimize the risk of a strategic exchange with the Soviet Unio~ with th
0107 153 man gives food, care and protection in exchange for the different services the
0037 153 Flanders. The printer was entranced to exchange a few of the place-names which
0028 152 when its shares are introduced to the Exchange, probably in January. He also
0025 152 to freer, cheaper and more widespread exchange of information between the ric
0123 152 ne was out of order; on the second the exchange was closed for a religious hol
0125 151 ng to hand back all their conquests in exchange for the tiny border enclaves

A similar range of good collocates and phrases is shown here, although both lines 11 and 66, with the unlikely collocates 'unbearable' and 'Unpleasantries', are far from typical, in terms of their lexical content, although they do exhibit typical grammatical constructions.

8.2.4. arts-prons/1/open/raw

12,404 total bonds, 176 bonded lines, SD: 29.6

0092 135 he flat and go to a hotel. The rate of exchange in Denmark is heavily against
0007 133 endeavoured to heal the wounds. In an exchange of letters with Mansholt he de
0071 133 ds over her identity to her husband in exchange for a small portion of his, sh
0052 132 amental right to adequate treatment in exchange for being deprived of his libe
0163 130 tinued to a imported, and there was an exchange of "light" North Sea oil for "
0086 127 ght to harvest any farmer's fields. In exchange, they get one ninth of the cro
0073 125 e disposed of, and offered for sale in exchange for cash - and when cash is no
0099 119 inder of the session was devoted to an exchange on the compatibility of religi
0067 118 d raw materials naturally need foreign exchange to buy these, but because of t
0153 117 s of bacteria and mammalian gnes may exchange in nature. So it may be that
0136 116 ound 80. Bear in mind that the rate of exchange while 1 was there was 11.20 fr
0090 116 hat is, to establish the exact rate of exchange at which mechanical energy is
0018 114 of value, the proposition that things exchange in accordance with the amount
0123 114 ne was out of order; on the second the exchange was closed for a religious hol
0109 114 mber of units in any of our Trusts in exchange for your securities - this exc
0084 114 for the walls of their private bars in exchange for a few pints of beer. Or, i
0143 113 pt to minimize the risk of a strategic exchange with the Soviet Unio~ with th
0167 112 ush to sell. This hit the floor of the Exchange with torrential force. The mac
0025 108 to freer, cheaper and more widespread exchange of information between the ric
0145 107 re businessmen of various sorts met to exchange money, property or goods. Afte
0064 107 conomy changed from one based upon the exchange of goods and services to one b
0037 107 Flanders. The printer was entranced to exchange a few of the place-names which
0132 105 ome s"pose" and so on. There is a sure exchange :CHANGED of thought and some p
0054 105 and manufactures to the third world in exchange for raw materials and food, is

```
0154 104 sion. Another sent to me by the Labour Exchange presumably out of sheer kindne
```

In this set of output, which differs from the previous sets in that it was created using a non-empty stopword list, it is interesting that the grammatical constructions are still heavily represented, although the 'in exchange for' phrase is now present more frequently than in the earlier sets. This is undoubtably due to the suppression of the 'the/an exchange' link, introduced by the use of the 'arts-prons' stopword list, since this construction occurs most often in the company of a strong lexical item, as in line 64, 'the exchange of goods...'. As might be expected from the use of a non-empty stopword list, the number of bonds formed tails off rather more quickly than with the 'zero' stopword list and the total number of bonds is correspondingly lower.

8.2.5. zero/2/open/raw

9,208 total bonds, 176 bonded lines, SD: 29.4

```
0092 134 he flat and go to a hotel. The rate of exchange in Denmark is heavily against
0152 116 rovides a national focal point for the exchange of information, ideas and expe
0084 114 for the walls of their private bars in exchange for a few pints of beer. Or, i
0060 109 ate a small plot on the worst land, in exchange for agricultural and even dome
0044 107 a and are the same kind of judgemental exchange these ladies, as children, ove
0025 105  to freer, cheaper and more widespread exchange of information between the ric
0007 105  endeavoured to heal the wounds. In an exchange of letters with Mansholt he de
0064 105 conomy changed from one based upon the exchange of goods and services to one b
0054 105 and manufactures to the third world in exchange for raw materials and food, is
0123 102 ne was out of order; on the second the exchange was closed for a religious hol
0066 101   d an expert in Round Tableip and the Exchange of Unpleasantries. Last certai
0163 100 tinued to a imported, and there was an exchange of "light" North Sea oil for "
0143 099 pt to minimize the risk of a strategic exchange with the Soviet Unio~ with th
0037 098 Flanders. The printer was entranced to exchange a few of the place-names which
0161 096 the night officer and the sister would exchange a few words with us. In my fir
0011 096  here was unbearable. And he wanted to exchange the unbearable for the very ba
0086 094 ght to harvest any farmer's fields. In exchange, they get one ninth of the cro
0150 093 rld have a chance to meet together and exchange ideas. The Vegetarian Federal
0090 093 hat is, to establish the exact rate of exchange at which mechanical energy is
0099 092 inder of the session was devoted to an exchange on the compatibility of religi
0006 091  could utter silence. In practice, the exchange of letters often takes a full
0146 090 re of the enormous destruction such an exchange would cause, and this awarenes
0172 086  wine later, and frolicketwith an oral exchange of that, laughing over the hyd
0138 083 ouse production a main item of foreign exchange or moneyearning, only the simp
```

```
0097 083 id enjoyment. The justification is the exchange of ideas, and the value of thi
```

This set of output is the first to be presented here which uses a link threshold higter than one. This causes an overall lower degree of bonding and highlights lines which contain strong grammatical features, especially the 'in exchange for' and 'the/an exchange of'.

8.2.6. zero/1/fixed/rel

10,732 total bonds, 176 bonded lines, SD: 28.3

```
0007 121  endeavoured to heal the wounds. In an exchange of letters with Mansholt he de
0140 115 papers presented to the Securities and Exchange Com- mission, the multinationa
0097 115 id enjoyment. The justification is the exchange of ideas, and the value of thi
0038 112 Lou Darrow Carrington runs the foreign exchange desk for the bank's corporate
0087 107 gned to work with the local service to exchange information, to train the loca
0006 106  could utter silence. In practice, the exchange of letters often takes a full
0002 106  Ruder, chairman of the Securities and Exchange Commission, said Britain, the
0060 105 ate a small plot on the worst land, in exchange for agricultural and even dome
0027 104  two marine insurance firms, the Royal Exchange Insurance and the London Assur
0055 104 and other statues from the first Royal Exchange). The Library is a first-class
0054 104 and manufactures to the third world in exchange for raw materials and food, is
0167 103 ush to sell. This hit the floor of the Exchange with torrential force. The mac
0018 103  of value, the proposition that things exchange in accordance with the amount
0079 102 eneral impression. He was moved by the exchange of vows, the old clear words,
0130 100 oco had reported to the Securities and Exchange Commission. Recently, Texaco l
0015 100  know where they gave the best rate of exchange. The whole place was reflected
0029 099  which we have witnessed on the stock exchange this week, does the team agree
0161 099 the night officer and the sister would exchange a few words with us. In my fir
0173 098 y another name. In the absence of real exchange controls, however, the tax aut
0102 098 ld be impossible for the "fighters" to exchange roles so freely. They would be
0152 097 rovides a national focal point for the exchange of information, ideas and expe
0101 097 kbrokers did by the pillars of the old Exchange. Through the west gate of this
0028 095  when its shares are introduced to the Exchange, probably in January. He also
0123 095 ne was out of order; on the second the exchange was closed for a religious hol
0064 095 conomy changed from one based upon the exchange of goods and services to one b
```

This set contains lines with strong lexical collocates – 'information', 'vows', while also featuring the grammatical patterns seen previously.

8.2.7. arts-prons/1/open/rel

8,164 total bonds, 176 bonded lines, SD: 23.1

```
0084 106 for the walls of their private bars in exchange for a few pints of beer. Or, i
0099 104 inder of the session was devoted to an exchange on the compatibility of religi
0007 103  endeavoured to heal the wounds. In an exchange of letters with Mansholt he de
0071 101 ds over her identity to her husband in exchange for a small portion of his, sh
0052 101 amental right to adequate treatment in exchange for being deprived of his libe
0163 099 tinued to a imported, and there was an exchange of "light" North Sea oil for "
0086 095 ght to harvest any farmer's fields. In exchange, they get one ninth of the cro
0073 091 e disposed of, and offered for sale in exchange for cash - and when cash is no
0090 088 hat is, to establish the exact rate of exchange at which mechanical energy is
0092 085 he flat and go to a hotel. The rate of exchange in Denmark is heavily against
0143 083 pt to minimize the risk of a strategic exchange with the Soviet Unio~ with th
0054 082 and manufactures to the third world in exchange for raw materials and food, is
0125 081 ng to hand back all their conquests in exchange for the tiny border enclaves
0167 080 ush to sell. This hit the floor of the Exchange with torrential force. The mac
0039 080 MALA TO GIVE UP ITS CLAIM TO BELIZE IN EXCHANGE FOR CERTAIN ECONOMIC CONCESSIO
0037 080 Flanders. The printer was entranced to exchange a few of the place-names which
0154 079 sion. Another sent to me by the Labour Exchange presumably out of sheer kindne
0050 079 all of us if we did not calm down. Our exchange was heated. Within a matter of
0025 078  to freer, cheaper and more widespread exchange of information between the ric
0123 077 ne was out of order; on the second the exchange was closed for a religious hol
0014 077  it ahead of her, how it would be. The exchange of witty letters, fewer as tim
0024 075  them (as used to be said on the Stock Exchange), cast no doubt envious glance
0145 075 re businessmen of various sorts met to exchange money, property or goods. Afte
0136 073 ound 80. Bear in mind that the rate of exchange while 1 was there was 11.20 fr
0109 073  mber of units in any of our Trusts in exchange for your securities - this exc
```

The combination of 'arts-prons' stopwords with the relative link type forefronts the 'in exchange for' construction.

8.2.8. btb/1/open/raw

4,618 total bonds, 172 bonded lines, SD: 21.9

```
0060 080 ate a small plot on the worst land, in exchange for agricultural and even dome
0036 076 Copies were burned on the London Stock Exchange and destroyed at exchanges in
0009 074  foreign services usually press for an exchange, and often in poor countries t
0129 072 ocal shop. We can give you sterling in exchange for most foreign notes but coi
0084 067 for the walls of their private bars in exchange for a few pints of beer. Or, i
0045 067 a for them, and I keep him supplied in exchange for plenty fires and troubles
0109 065  mber of units in any of our Trusts in exchange for your securities - this exc
0125 064 ng to hand back all their conquests in exchange for the tiny border enclaves
0080 064 er was an indeterminate little man. In exchange for our money, they were suppo
```

Collier

0104 063 ltivation can be surprisingly high, in exchange for no investment in fertilize
0073 063 e disposed of, and offered for sale in exchange for cash - and when cash is no
0107 062 man gives food, care and protection in exchange for the different services the
0094 062 hey would be given decent treatment in exchange for "honest labor." ZOB iasued
0091 062 hat the offence might be overlooked in exchange for a consideration: they woul
0071 062 ds over her identity to her husband in exchange for a small portion of his, sh
0048 062 al - and end up with failed degrees in exchange for a phenomenal understanding
0039 062 MALA TO GIVE UP ITS CLAIM TO BELIZE IN EXCHANGE FOR CERTAIN ECONOMIC CONCESSIO
0032 062 , price controls and food subsidies in exchange for voluntary wage restraint,
0141 061 pawn your land for five years or so in exchange for the cash. The moneylender
0054 061 and manufactures to the third world in exchange for raw materials and food, is
0122 060 nd ack- nowledged him as their Lord in exchange for whatever they most desired
0052 060 amental right to adequate treatment in exchange for being deprived of his libe
0034 060 And priests were extracting "gifts" in exchange for burying non-churchgoers in
0123 058 ne was out of order; on the second the exchange was closed for a religious hol
0098 057 imise our clients' exposure in foreign exchange. We tell them what's happening

The use of the 'btb' stopword list, which contains no prepositions, boosts the 'in exchange for' pattern enormously in this set of output, almost to the exclusion of all other immediate collocational features. Only the presence of strong lexical items serves to differentiate the lines in this set: 'land', 'money', 'materials', although there is little variation across the bond scores in this output, suggesting that the heavily repeated grammatical features are largely responsible for the bonding.

8.2.9. arts-prons/1/fixed/raw

7,100 total bonds, 176 bonded lines, SD: 21.9

0052 110 amental right to adequate treatment in exchange for being deprived of his libe
0109 095 mber of units in any of our Trusts in exchange for your securities - this exc
0007 094 endeavoured to heal the wounds. In an exchange of letters with Mansholt he de
0099 091 inder of the session was devoted to an exchange on the compatibility of religi
0092 091 he flat and go to a hotel. The rate of exchange in Denmark is heavily against
0152 084 rovides a national focal point for the exchange of information, ideas and expe
0006 083 could utter silence. In practice, the exchange of letters often takes a full
0037 083 Flanders. The printer was entranced to exchange a few of the place-names which
0077 082 ehall but best remembered by me for an exchange of Jack Buchanan's signed ciga
0136 069 ound 80. Bear in mind that the rate of exchange while 1 was there was 11.20 fr
0071 069 ds over her identity to her husband in exchange for a small portion of his, sh
0166 068 unting, president of the Toronto Stock Exchange. Mr Walker acknowleged that t
0159 068 tein, but it would only precipitate an exchange of feelings on a subject which
0039 068 MALA TO GIVE UP ITS CLAIM TO BELIZE IN EXCHANGE FOR CERTAIN ECONOMIC CONCESSIO
0143 067 pt to minimize the risk of a strategic exchange with the Soviet Unio~ with th

```
0131 067 olas Goodison, ? Chairman of the Stock Exchange, was asked if he found the lar
0167 065 ush to sell. This hit the floor of the Exchange with torrential force. The mac
0112 065 means of production, distribution, and exchange is profitability; that any dep
0082 064 ferent workplaces and jobs can meet to exchange their experiences. In other wo
0138 063 ouse production a main item of foreign exchange or moneyearning, only the simp
0028 062  when its shares are introduced to the Exchange, probably in January. He also
0017 062  new house (the 10 per cent deposit on exchange of contracts) before you've re
0014 062  it ahead of her, how it would be. The exchange of witty letters, fewer as tim
0132 061 ome s"pose" and so on. There is a sure exchange :CHANGED of thought and some p
0048 061 al - and end up with failed degrees in exchange for a phenomenal understanding
```

This set of output shows a combination of lexical and grammatical features, similar to the earlier 'arts-prons' example.

8.2.10. zero/2/open/rel

5,282 total bonds, 176 bonded lines, SD: 20.3

```
0084 088 for the walls of their private bars in exchange for a few pints of beer. Or, i
0099 086 inder of the session was devoted to an exchange on the compatibility of religi
0054 086 and manufactures to the third world in exchange for raw materials and food, is
0040 080 N A SINGLE IMAGINATIVE GESTURE. AT THE EXCHANGE WE GET THE OLD SPECTACLE OF A
0092 078 he flat and go to a hotel. The rate of exchange in Denmark is heavily against
0037 078 Flanders. The printer was entranced to exchange a few of the place-names which
0143 076 pt to minimize the risk of a strategic exchange with the Soviet Unio~ with th
0068 069 d, "is not the way to begin a cultural exchange." The incident caused the trai
0060 067 ate a small plot on the worst land, in exchange for agricultural and even dome
0101 064 kbrokers did by the pillars of the old Exchange. Through the west gate of this
0097 064 id enjoyment. The justification is the exchange of ideas, and the value of thi
0007 064  endeavoured to heal the wounds. In an exchange of letters with Mansholt he de
0167 063 ush to sell. This hit the floor of the Exchange with torrential force. The mac
0066 062  d an expert in Round Tableip and the Exchange of Unpleasantries. Last certai
0142 060 pproached the defense table, hoping to exchange a few words with them. The gu
0055 060 and other statues from the first Royal Exchange). The Library is a first-class
0079 059 eneral impression. He was moved by the exchange of vows, the old clear words,
0140 058 papers presented to the Securities and Exchange Com- mission, the multinationa
0173 057 y another name. In the absence of real exchange controls, however, the tax aut
0123 057 ne was out of order; on the second the exchange was closed for a religious hol
0053 057 and maintain the correct rate of fluid exchange within the various fluid compa
0131 056 olas Goodison, ? Chairman of the Stock Exchange, was asked if he found the lar
0002 056  Ruder, chairman of the Securities and Exchange Commission, said Britain, the
0163 055 tinued to a imported, and there was an exchange of "light" North Sea oil for "
0152 054 rovides a national focal point for the exchange of information, ideas and expe
```

This final set of output, only the second set presented here with a link threshold of two, contains an interesting mix of grammatical patterns and strong collocational features. There is also a good range of bond scores across the lines and the total number of bonds remains quite high, while the standard deviation reflects a useful degree of differentiation across all 176 lines in the output.

8.3. Remarks

The samples from the output of *cohort* given above exemplify many contextual features of the node word 'exchange' which seem to correspond to (my) linguistic intuitions of what is typical and as such they are very encouraging. The most obvious grammatical patterns are 'in exchange for' and 'the/an exchange of X' and 'to exchange X' and these are seen most clearly in output where the majority of grammatical items have been suppressed, leaving only the prepositions and lexical items, most notably in the 'btb/1/open/raw' set.

There are numerous instances of strong (lexical) collocates ('rates', 'information', 'letters', 'foreign', 'Royal', 'money', 'cash' for example), all of which are intuitively satisfying, especially where these also occur in conjunction with the aforementioned grammatical constructions. Examples of this can be seen in lines such as 54: 'in exchange for raw materials and food', 152 'the exchange of information, ideas and expe[riences]' and 142 'to exchange a few words'.

Not all the results are entirely satisfactory – there are several lines present in the samples which seem somewhat unusual; lines 11, 37 & 66, containing the rather untypical collocates 'unbearable', 'entranced' and 'Unpleasantries' have scored quite highly in some of the sets of output. This can only be due to the grammatical constructions within the line, as these lexical items do not occur elsewhere and so cannot be responsible for the formation of links.

The presence of not-so-desirable lines in the top-scoring section of the output brings us to the difficult issue of the evaluation of the output from *cohort*. Deciding which of the many different sets of output is optimal is a difficult and subjective task and the analysis presented here is not intended to be anything more than an illustration of the capabilities of the software. What is required is a systematic evaluation exercise which is capable firstly of determining whether the results of the selection system are at all acceptable and secondly of identifying the combinations of parameters which produce the best results. In the chapter which follows, the methodology and results of such an exercise, carried out with the aid of a group of experienced corpus users, are presented.

Chapter 9

Evaluation

9. Evaluation of the Automatic System

9.1. Introduction

In order to determine whether the results produced by the automatic system correspond to the intuitions of corpus users, an evaluation exercise was carried out which involved comparing the lines selected by corpus users as representative with the scores applied by the system. This chapter will present the results of this comparison.

9.2. Need for Evaluation

The *cohort* software operates by identifying the links and bonds present between the lines which make up the concordance text. While the means by which the bonds are established is relatively simple, the large number of configurations made possible by varying different parameters creates a considerable task in terms of the validation of the results.

The software has been in use for some time now and appears to be (in software engineering terms) *correct*. That is, for a given set of input, it produces a matrix of interrelations between the elements of the input which corresponds to expectations, which is to say that it is the same as would be produced by a manual analysis, if this were to be limited to the forms of link allowed under the automatic system. Analysing a concordance automatically can be carried out much more quickly than manually and thus the possibility exists of producing many sets of output, each based on a different configuration of parameters. The point of this would be to attempt to establish which configuration created the best set of results.

The evaluation of a large number of outputs poses a problem however. It has already been shown that corpus users are not able to make full use of the information contained in concordance lines, since they are unable to retain an overview once the number of lines gets very large (more than about 1,000). It is therefore not feasible for them to evaluate the output from the automatic system, since this simply consists of concordance lines and

would therefore be no more easily tractable than concordance lines drawn straight from the corpus.

An alternative strategy would be to impose a numerical cutoff and ask the corpus users to evaluate only those lines which achieved a high bond score. The disadvantage of this approach is that it would disregard the fact that there may be 'good' lines which did not receive a high score. Consequently, *all* lines in all the sets of output would need to be examined (a one-off evaluation of the 200 lines would not be sufficient) since each configuration of parameters will theoretically identify different characteristics of the concordance. If every one of the 252 possible configurations of parameters were used, however, this would result in 2,825,424 lines to be examined for the word 'date' alone. As it is impracticable to verify this many concordance lines, it is likely that this strategy would bias the results in favour of the software, because the omission of 'good' lines from automatic selection would not be detectable.

Earlier, it was mentioned that the goal of this exercise was to establish which set of parameters produced the best results. This naturally raises the question of what is 'best'. The solution presented here is to compare the automatically-produced sets of output with human intuition. Naturally, it would be preferable to have access to many concordance users' intuitions about many sets of output for several different node words. Unfortunately, however, it is not practicable for large numbers of experienced corpus users to inspect and evaluate all the various output options and so a less time-consuming method had to be employed.

The evaluation process was complicated further by the disappearance of most of the respondents, following staff cutbacks at Cobuild. This has made it impossible to use any other evaluation methods as a means of comparison or to repeat the evaluation using different data. Another alternative might have been to compare the results obtained from using a different analysis tool, *typical*, mentioned in Section 5.2, but this is not publicly available and, in contrast to *cohort*, is dependent on information derived from the source

corpus of the concordances, making it difficult to use outside of the Cobuild context, which is where the concordances were generated. No other tool exists which has similar functionality to typical and *cohort*.

9.3. Scope of the Evaluation

The material used in the evaluation consisted of a random selection of 200 concordance lines for the word 'date', extracted from the BoE, which at the time contained approximately 16,000 occurrences of the word. 'Date' was chosen because it occurs frequently (16 times more so than the 1,000 occurrences which represents 'too much' for most corpus analysts interviewed); it crosses word classes (noun, verb) and is also mildly polysemous within each word class. As such, it was thought to be a useful test of the capabilities of the analytical software. In addition, 'date' has a number of strong collocational patterns, which, it is to be hoped, will be identified by the software.

Twelve Cobuild lexicographers and grammarians were asked to select from the 200 lines up to twenty which they considered to be representative of the usage of the word. As a secondary exercise, they then had to choose up to twenty lines which they considered most usable as examples in one of their reference texts. The definitions of 'representative' and 'usable', as presented to the respondents, are as shown in sections 9.4.1 and 9.4.2 respectively.

It was noted in Section 2.5.2.2 that many of the factors influencing the usability of a concordance line (as an example in a reference work) were external or meta-textual, real world references, contentious issues etc. There should therefore be no means by which *cohort* could identify usable lines. The usability data was nevertheless collected from the respondents, in order to test whether the cohesive patterns of the concordance might in some way correlate with usability. Since the issue of usability is a secondary one, intermediate results will not be presented for all the tests involving a comparison of the output of *cohort* with the lines selected as usable, but these can be found in summary in the final

section of this chapter. A full set of results is presented for the comparisons of automatic and manual, representative lines, however.

No time restriction was placed on the selection exercise and the corpus users were encouraged to carry it out at whatever speed represented their normal working methods. It was hoped that this approach would provide the most realistic setting for the selection process and so produce more authentic results. In order to obtain an overall impression of which concordance lines were preferred, the selections from each of the respondents were collated. This method allowed a score to be assigned to each line on the basis of the number of informants who chose it; that is, if four respondents considered a particular line to be representative, then it received a score of four. The advantage of this method was that each respondent was only presented with 200 lines to evaluate, albeit according to the two different criteria of representativeness and usability, and had no contact with the results of the automatic system. It is conceivable that, had they received lines which had been pre-analysed by the *cohort* software, the informants' judgement might have been influenced, since, as seen in Figure 5.6, the automatic analysis adds extra information to the concordance, resulting in a slightly different format. The bond score attached by the software might also have influenced their selection, as they may have been tempted to choose lines in accordance with or even counter to the analysis made by the software.

The automatic system was run over the same 200 lines using several different sets of parameters, producing nine different rankings, based on the number of bonds each line attained. A Pearson's Correlation Coefficient score was then calculated for each of the automatically ranked sets, comparing the number of bonds identified by the system with the scores derived from the corpus users' selections.

9.4. The Concordance Questionnaire

The evaluation was carried out using a combination of methods, one hardcopy and one electronic, within the framework of a simple questionnaire which firstly asked the

respondents to choose concordance lines which they felt to be representative or usable as examples and secondly required them to introspect on the mental processes which they performed in order to make the selection.

Prior to the evaluation, the corpus users were canvassed as to their preference for the format of presentation of the concordance evidence. Most responded that they would prefer to have online access to the corpus data, as this would be most similar to their everyday experience of using the corpus. Given that this approach would be expected to yield the most realistic results, a feature of the Cobuild corpus software system was exploited which enabled each user to be presented with the same random set of lines. It was thus possible to provide all the usual corpus analysis tools to those who wished to make use of them. For those corpus users who were content to work with hardcopy data, the same set of lines was printed out using various sort options: -1, -2 and +1, relative to the node word. The 'columns' and 'collocates' analyses (see Section 1.3 for details) of the 200 lines were also provided.

9.4.1. Selection of Representative Lines

The respondents were set the following task:

Examine the citations and select the twenty lines which you think are most representative of the behaviour of the node word. Feel free to make use of all the versions of the data. If you are unable to identify twenty lines, select as many as you think are representative.

Please enter the numbers of the citations you select in the boxes below. You do not need to rank the citations, so the order in which you enter them is not important.

Twenty boxes were then provided in which the respondents could enter the numbers of those lines which they held to be representative. Each of the votes cast by means of these boxes was added to the total for each line.

9.4.1.1. Summary of Results

The twelve informants made a total of 211 selections. The shortfall from the possible maximum of 240 selections (12 × 20) is accounted for by the fact that six respondents were not able to identify twenty representative citations. The responses can be summarised as:

Total votes cast: 211
Number of lines selected: 92

Maximum votes for any one line: 9 (1 case)
Minimum votes for any one line: 1 (39 cases)

50th centile (actually 50.2) of votes accounted for by 23 lines (26%).
75th centile (actually 74) of votes accounted for by 9 lines (10%).

From this summary it can be seen that a quarter of the lines (23 out of 92) received over half the votes (106 out of 211). The ratio is higher still if one examines the values above the 75th centile, the top-scoring nine lines, which account for over a quarter of the votes. This can be expressed as percentages as follows: (55 : 211) vs (9 : 92) = 26% vs 10%, a ratio of two and a half to one.

It should be noted that the centile calculations have been rounded to 50 and 75 for the purposes of comparison with the results from the 'usable' section which follows. In the 75th centile set, for example, line 174 should not be excluded on the basis of its ranking in the list; it is simply the last item in a range of lines which achieved a score of 5.

9.4.1.2. Results in Detail

The scores for each manually-selected line are presented in columns in Table 9.1, ranked in descending order of score. The 'S' columns give the score and the 'L' columns give the concordance line number. The actual concordance lines selected by the respondents can be found in Appendix 4.

S	L	S	L	S	L	S	L	S	L	S	L	S	L	S	L	S	L
9	159	4	30	3	125	2	28	2	115	2	198	1	33	1	107	1	141
7	57	4	75	3	133	2	35	2	116	2	199	1	38	1	110	1	142
6	4	4	84	3	137	2	52	2	124	1	5	1	40	1	119	1	144
6	6	4	121	3	172	2	54	2	126	1	7	1	49	1	122	1	155
6	55	4	151	3	190	2	56	2	129	1	9	1	58	1	127	1	169
6	143	4	162	3	200	2	67	2	132	1	11	1	68	1	128	1	178
5	25	3	23	2	21	2	98	2	154	1	13	1	72	1	130	1	180
5	65	3	42	2	22	2	105	2	164	1	19	1	82	1	131	1	183
5	87	3	77	2	24	2	106	2	188	1	27	1	83	1	135	1	187
5	174	3	96	2	26	2	108	2	191	1	29	1	97	1	136	1	189
4	10	3	123														

Table 9.1: Manual Scores for Representative lines

9.4.1.3. Conclusions

What we find, then, is that as far as a few lines are concerned, there is a high degree of agreement among the respondents as to which lines are the most representative. The somewhat lower scores further down the ranking suggest that there was rather less agreement, since 27 out of the 92 lines selected (29%) scored two, while 39 lines (42%) were only chosen by one informant.

9.4.2. Selection of Usable Lines

The respondents were asked to do the following:

Re-examine the citations and select twenty which you would be feel would be suitable for use as examples in a dictionary. You may assume that the citations could be edited to some extent, that is, it is possible that only a part of a citation would be used, or that the citation would be expanded to a full sentence. As above, if you feel that there are not twenty usable examples, you may select fewer.

The citations you choose here do not have to overlap with the lines you selected in the previous section, but it does not matter if they do. Please enter the numbers of your selected lines into the boxes below. Here too, order is not important.

Collier

A further twenty boxes were provided for the numbers of those lines thought to be most usable.

9.4.2.1. Summary of Results

When asked to select concordance lines on the basis of usability, only four of the respondents chose a full twenty. This reflects the issues relating to the suitability of concordance lines which were raised in Chapter 2 and contrasts with the results in the 'representative' section above, where six corpus users felt able to select twenty lines.

The reason for the difference is probably that it is easily possible for a concordance line to contain one or more regular features of the node word which might make it a representative line, but there are many textual and non-textual criteria to be applied before it can be said to be a usable line.

Total votes cast: 191
Number of lines selected: 82

Maximum votes for any one line: 8 (3 cases)
Minimum votes for any one line: 1 (33 cases)

50th centile (actually 50.5) of votes accounted for by 20 lines (25%).
75th centile (actually 74.9) of votes accounted for by 7 lines (9%).

9.4.2.2. Results in Detail

The scores for each line are given in the columns below, ranked in descending order of score:

S	L	S	L	S	L	S	L	S	L	S	L	S	L	S	L
8	57	4	58	3	183	2	75	2	172	1	23	1	83	1	129
8	143	4	115	2	13	2	108	2	178	1	27	1	85	1	132
8	159	4	133	2	19	2	116	2	189	1	28	1	89	1	135
7	162	4	198	2	24	2	121	2	190	1	34	1	96	1	142
7	174	3	4	2	32	2	125	2	191	1	38	1	105	1	155
5	55	3	10	2	33	2	136	2	199	1	49	1	107	1	160
5	65	3	26	2	35	2	137	2	200	1	56	1	109	1	161
5	84	3	30	2	42	2	147	1	1	1	61	1	110	1	169
5	87	3	76	2	52	2	154	1	9	1	74	1	123	1	187
4	6	3	77	2	67	2	164	1	11	1	79	1	126	1	188
4	25	3	151												

Table 9.2: Scores for Usable lines

9.4.2.3. Conclusions

As was the case with the scores from the 'representative' analysis, a rapid fall-off of scores is noticeable, but it is slightly less marked than it was in Table 9.1. In this set, 50% of the votes were cast for only 20 lines, with slightly more lines being selected twice (26 or 32% as opposed to 29%) and marginally fewer selected just once: 33 (40%) compared with 42% previously.

9.5. Additional Questions

Since the input to this exercise consisted of randomly sampled lines, it was felt that this would be an ideal opportunity to collect more information on various aspects of using this type of data. As was seen in the chapter on using large corpora, random sampling is quite often used by users of large textual databases to assist them in dealing with the huge amount of evidence with which they are confronted. As noted previously, the major drawback of the random sample is that it creates the possibility that some feature of the node word will be missed by the corpus analyst, simply because it is not included in the sample. The first set of additional questions therefore attempts to establish whether this happened in this instance.

9.5.1. Section i)

The data with which you were provided was only a sample of the occurrences of the word 'date', of which there are over 16,000 in total. Do you feel that the sample adequately represented the characteristics of the node word?

Only one out of the twelve respondents replied unequivocally yes to this question. All others pointed out that the verb usage was severely under-represented.

You may now check your intuitions against the corpus if you wish. Were they correct?

All of those who did check against the full corpus confirmed the general consensus that the verb uses were not sufficiently represented. In fact, there are 3,066 instances of 'date' as a verb, which is a considerable proportion of the 16,818 total occurrences, yet this sample, albeit a rather small one, all but ignored this usage, with only eight lines involving 'date' used as a verb: 4% for the sample of 200 lines as opposed to 19% in the corpus as a whole.

What size sample would you have chosen for an initial examination of this node word and why would you choose this size?

Responses here ranged from '30 to 50' up to as much as 2,000, but several respondents drew attention to the fact that they would choose to examine the node word in the context of all the forms of its lemma, for example, selecting 100 lines per part of the verb.

9.5.2. Section ii)

How many senses of the node word were you able to identify from the 200 citations?

Here too there was a wide range of answers. One respondent identified as many as eighteen senses, another as few as four. This is firstly attributable to the interpretation of the word 'sense'. Those corpus users who gave a high figure were possibly thinking in terms of dictionary entries or sub-entries, so that 'up to date' would be included as a separate 'sense' from the main 'specific point in time' sense.

The second factor in the number of senses identified has to do with the characteristics of the corpus analyst. It is generally stated (at least among Cobuild corpus users) that there are two basic types of concordance analysts: *lumpers* and *splitters*. As the names suggest, the former are more likely to identify fewer senses in a given set of concordance lines than the latter. The variation in the number of senses found in the sample set suggests that the respondents included both lumpers and splitters.

Briefly list which senses you identified.

The senses identified by more than one respondent were:

- A point in time
- A romantic meeting, the person that you romantically meet, also the verb
- To identify the time of an event
- A pre-arranged event
- A performance (e.g. on a concert tour)
- The phrase 'to date', meaning 'until now'
- To 'date back to' or 'date from' a point in history
- The fruit
- To mark the date on something

Interestingly, the sense of 'go out of fashion' is not included by a single respondent. This is for the simple reason that this sense is not present in the sample set – further proof of the shortcomings of the random selection.

Do you feel that these are all the senses of this node word?

Only two of the respondents were able to agree unequivocally that the senses shown in the sample represented all the possible ones for this node word. Some referred loosely to the lack of verb senses, while one or two gave additional senses, specifically 'go out of fashion' and 'show you to be older than the people around you'.

9.5.3. Section iii)

When you are beginning your analysis of a word (any word), on what grounds (e.g. total frequency, number of senses expected, distribution across sub-corpora) do you choose the number of sample lines to examine?

There was great variety in the responses to this question. They are therefore presented separately, albeit in summarised form. The bracketed figures following each response identify which informant made the response.

'If there are less than 500 total occurrences of the node word, look at them all, otherwise take a 500 line sample.' (11)

'Depends on: the number of senses, the total frequency and the number of grammatical possibilities expected; an unexpected distribution across the sub-corpora or surprising collocates would be grounds for closer scrutiny.' (1)

'100 line sample regardless, as an initial examination; for including or dropping a given sense, use the whole corpus + search/picture etc.' (2)

'If there are less than 2,000 occurrences, look at them all, otherwise start with 10% of the lines.' (5)

'Up to 100 occurrences, look at them all; for 100-500, take 100; for more than 1,000, use 200.' (8)

'For words with higher frequency or more senses, take a larger sample.' (4)

'The number of senses/patterns/collocates expected.' (3)

'Take 10% sample unless word is very high or low frequency. Also depends on lexicographical task in hand: editing existing text or compiling new.' (6)

'Small initial samples to identify senses and syntax, followed by more detailed

analysis of collocates, lemma and sub-corpus distribution.' (10)

'Frequency and number of senses expected.' (9)

It can be seen that a wide range of criteria are applied by corpus users as they set about the analysis of a set of concordance lines.

9.6. General Comments on the Manual Analyses

The 200 lines analysed by the corpus users and processed by the automatic system represented only a 1.3% sample of the more than 16,000 total occurrences of the node word 'date'. It is interesting to note that of the 200 concordance lines which the respondents could have chosen, only 105 lines across the two sets of results were actually selected, corresponding to 44% (average ±3%) of all possible lines. While this figure only relates to a small random sample of one word, a sample which has already been shown to be under-representative, it seems worth speculating as to whether it is illustrating a more generally-applicable principle. This principle would state that only a limited proportion of the contexts of a node word contain enough features for them to be recognised by corpus users as being representative of that node's environment. The corollary of this would be that there are occurrences of a node word, a considerable proportion in fact, which are far from typical. This is confirmed by the opinion of many Cobuild corpus users that a substantial part of the evidence which the corpus supplies is unusable, on the grounds that it is not representative. In a later section, some alternatives to comparison with intuition are explored and an analysis is carried out which attempts to underpin this principle with evidence from the full set of concordance lines for 'date'.

As was illustrated in Chapter 2, those features of a concordance line which make it suitable for use as an example in a reference work tend to be concrete, surface characteristics such as strong collocational patterns or syntax. We can take it for granted that a line which is judged to be 'usable' is also 'representative', although the converse does not apply, a representative line being subject to examination for the extra-textual 'negative'

features such as offensiveness or cultural-specificity, or the presence of rare lexical items, before it can be deemed usable. Confirmation of this is to be found in the figures for the number of lines selected by the respondents for the two criteria: 211 votes for 92 lines were cast for the representative group, but when it came to choosing usable lines, only 191 votes for 82 lines were cast. The fact that only 105 lines out of the 200 received any votes implies a considerable overlap between the two sets, suggesting that the respondents were being more strict in the selection of usable examples. This can be confirmed by examination of the selection data, which indicate that of the 82 lines rated as usable, 69 were also regarded as representative.

9.7. Comparison of Manual and Automatic Analyses

We shall begin with an examination of the number of lines which are selected by the program when it is run on the concordance for 'date' using various sets of parameters. In the table which follows, the parameters are listed in roughly ascending order of strictness, that is, the number of lines which form bonds decreases fairly consistently. These tests were run using the 'btb' stopword list† and did not impose a span restriction on the links, i.e. they were free to occur at any distance from the node word within the bounds of the concordance line. The stopword list and span represent two of the parameters to the automatic selection process and they remained relatively fixed (with only the substitution of one stopword list for another to be considered). Two other parameters, however, the link threshold and the link type were varied to a greater degree. For this initial comparison, link thresholds of 1, 2 and 3 were tried, and these are represented by the numbers in the 'Link Threshold' column of the table. The various link types (raw, relative and absolute) were used and these are shown under 'Link Type' as 'raw', 'rel' and 'abs' respectively.

† See Chapter 6 for the specifications of the different stopword lists.

Link Threshold	Link Type	No. of lines Selected
1	raw	199
1	rel	197
1	abs	188
2	raw	150
2	rel	150
2	abs	38
3	raw	46
3	rel	41
3	abs	9

Table 9.3: Number of lines selected using various parameters

It was noted above that only 105 lines out of the possible 200 were chosen by the respondents as either usable or representative. This contrasts with the results for one link shown here, where nearly all the lines are selected, suggesting that the majority of concordance lines have at least one significant feature. This is corroborated by evidence from the full set of 16,818 concordances for 'date', provided in the section 'Alternative Evaluation Methods' below.

As the first stage in the comparison of the automatic and manual results, a simple matching process can be performed on the lines selected by the two different methods. The two tables which follow show how many lines each of the automatic selections had in common with the two manual analyses. The entries are ranked in descending order of the number of common lines.

For the 'representative' comparison, the maximum possible number of common lines is 92 and for the 'usable' test it is 82.

Links	Type	common
1	raw	92
1	rel	92
1	abs	88
2	rel	72
2	raw	68
3	raw	26
2	abs	22
3	rel	21
3	abs	5

Table 9.4: Common lines between representative and automatic

Links	Type	common
1	raw	82
1	rel	81
1	abs	80
2	rel	66
2	raw	64
3	raw	27
3	rel	21
2	abs	20
3	abs	3

Table 9.5: Common lines between usable and automatic

The above figures can be easily misinterpreted, however. The score of 92/92 for one raw link in the 'representative' test fails to convey the fact that all but one line was present in that particular set of output. It is for this reason that only a small subset of the possible parameter permutations have been exemplified here.

For the full analysis of all the possible sets of output, a more rigorous method of comparison is required, one which takes into account the *negative* correlation as well as the positive. One suitable statistical test is the Pearson's Correlation Coefficient, or Pearson's Product-moment correlation coefficient. In this test, the scores for the manual and automatic analyses *including zero scores* are compared side by side. The mean score is

deducted from each one and the results are then multiplied. Where strong positive scores co-occur, a large positive product will result; similarly if strong negative scores co-occur again a strong *positive* product is produced. Where one score is negative and the other positive, the result is a negative product. An overall score is produced by summing the products and then scaling them to lie between -1 (strong negative correlation) and 1 (strong positive correlation). A score close to zero indicates little or no correlation. The tables below present a sample of the results of this test when used to compare each automatic analysis with the two manual ones. The items are presented in descending order of correlation as denoted by the 'r' score. The first table shows the ten parameter combinations which achieved the highest Pearson scores.

Stopword List	Link Threshold	Link Type	Compared with	r
zero	4	Raw	Repr	0.160472
zero	5	Raw	Repr	0.158715
arts-prons	3	Raw	Usable	0.151627
zero	3	Raw	Repr	0.140843
arts-prons	4	Raw	Usable	0.139588
arts-prons	2	Raw	Usable	0.137518
zero	2	Raw	Repr	0.137459
zero	3	Abs	Repr	0.134821
zero	3	Abs	Usable	0.133738
btb	2	Rel	Usable	0.13083

Table 9.6a
Pearson's Correlation Coefficient (r) for Automatic vs Manual Analyses
Top Ten Items

144 lines are omitted here. A table containing all the items from which the three lists shown here have been sampled can be found in Appendix 5a. The next table is drawn from the middle of the range of scores, which was largely occupied by combinations which showed no correlation, that is, received a score of 0.0. There were 44 such combinations.

Stopword List	Link Threshold	Link Type	Compared with	r
bt	6	Raw	Repr	0.0
bt	6	Raw	Usable	0.0
bt	6	Abs	Repr	0.0
bt	6	Abs	Usable	0.0
bt	6	Rel	Repr	0.0
bt	6	Rel	Usable	0.0
btb	6	Raw	Repr	0.0
btb	6	Raw	Usable	0.0
btb	6	Abs	Repr	0.0
btb	6	Abs	Usable	0.0

Table 9.6b
Pearson's Correlation Coefficient (r) for Automatic vs Manual Analyses
Mid-range Items

78 lines are omitted here, leaving ten combinations which displayed a weak negative correlation:

Stopword List	Link Threshold	Link Type	Compared with	r
bt	4	Abs	Repr	-0.0667674
bt	5	Abs	Repr	-0.0667674
bt	5	Rel	Repr	-0.0667674
btb	4	Abs	Repr	-0.0667674
btb	5	Abs	Repr	-0.0667674
btb	5	Rel	Repr	-0.0667674
top100	4	Abs	Repr	-0.0667674
top150	4	Abs	Repr	-0.0667674
top50	4	Abs	Repr	-0.0667674
zero	6	Abs	Repr	-0.0667674

Table 9.6c
Pearson's Correlation Coefficient (r) for Automatic vs Manual Analyses
Bottom Ten Items

For 200 objects, a statistically significant value of r would be greater than 0.14 or less than -0.14 †. What this indicates then is that there is almost no statistically significant

† I am grateful to Dr Paul Davies of the University of Birmingham for his advice. His derivation of the ± 0.14 value is as follows:

> The usual method for n (the number of objects)>30 is to use the fact that a transformed correlation coefficient has an approximate Normal distribution. If r is the correlation, calculate the 'z transform'
>
> z = 0.5*log{ (1+r)/(1-r)} and
> s = square root of (1/(n-2)).
>
> The quantity t = z/s has approximately the standardised Normal distribution if there is no true population correlation. So in the usual way there is evidence of significant correlation at 5% level if t lies outside the range (-2,2) and at 1% level if outside the range (-2.58,2.58). Note that the log is Napierian or hyperbolic to the base 'e' not log to the base 10 [which] means that a correlation r > 0.14 or r < -0.14 would be significant at 5% for n=200. For 1% probability the

correlation between either of the manual analyses and any of the automatic analyses, although certain sets of automatically selected lines do have a stronger correlation than others.

The lack of significant correlation between the manual and automatic analyses would at first glance appear rather disappointing. There are however two major factors to be considered when attempting to assess the worth of the automatically-selected sets of lines.

9.7.1. Correlation between Manual Analyses

On the whole, the manual analyses displayed very little consistency in the lines selected, although some sets correlated more closely than others, One way of visualising this is to create a matrix showing how many items were common to the 12 sets of selected lines. In the figure which follows, such a matrix is presented. The numbers in bold at the left and bottom represent the number of the set. The first 'cell' therefore compares set two with set one, conveying the information that they have three items in common.

```
 2  3
 3  3  3
 4  2  3  2
 5  4  3  2  1
 6  1  5  5  3  2
 7  3  4  3  4  7   6
 8  0  6  1  2  1   9  4
 9  0  6  2  3  2   6  5  5
10  1  6  3  3  4  10  8  8  7
11  1  2  3  2  3   3  4  6  3   4
12  0  8  2  6  4   7  5  7  7   7   3
    1  2  3  4   5  6  7  8  9  10  11
```

Figure 9.1: Common Items for Representative Lines

Some reasonably high values do appear in the matrix, but the level of overall correlation is not high. It is also worth considering that even if the lines had been selected completely randomly there would still have been some overlap, since it is impossible to choose twelve sets of twenty items out of 200 without some items being present in more than one

value would be 0.18. Of course these would be statistically significant but unlikely to be of any practical significance as the correlations are small.

set.

A similar situation exists for the lines which were rated as usable. The matrix of shared items looks like this:

```
 2  7
 3  0  2
 4  3  4  2
 5  3  3  2  5
 6  3  6  3  6  4
 7  3  6  4  6  7  6
 8  0  5  1  4  3  3  4
 9  3  6  1  3  2  4  3  3
10  3  6  3  8  4  9  6  3  5
11  3  6  3  4  4  6  7  6  4   6
12  0  2  1  3  3  3  4  2  2   4   2
    1  2  3  4  5  6  7  8  9  10  11
```

Figure 9.2: Common Items for Usable Lines

Again, there are a few pairings which do demonstrate quite a high degree of overlap, but overall there is little consistency.

The two matrices above are concerned with looking at lines which are shared across each possible pair of manually-selected sets of lines. This type of analysis presents a problem which is similar to the comparison of the automatic analyses with the manual ones in that any negative correlation is not taken into account. In order to perform the test more rigorously therefore, the manual selections were processed using vector analysis methods, resulting in a dissimilarity matrix for each set (representative and usable). This matrix contains a cell for each possible pairing of the manual analyses. Each cell contains a score, indicating the degree of dissimilarity between the pair of items. This score lies in the range of '0' to '1', with '1' representing the strongest possible dissimilarity and '0' being equivalent to no dissimilarity. The significance of intermediate values is hard to define as it depends on the type of data being compared, but in this instance any score higher than 0.78 is indicative of a strong dissimilarity. To illustrate how this works, if the two lists (1 3 5 7 9) and (2 4 6 8 10) were compared by this method, a single cell would result (since there is only one pair of lists to be compared) containing the value '1', as

there are no values in common. If the two lists were identical, however, the score would be zero, signifying minimum dissimilarity.

The method used is known as vectorisation and involves representing each list in a consistent format called a vector. If we call the two lists above A and B, the following vectors would be produced from them:

	1	2	3	4	5	6	7	8	9	10
A	1	0	1	0	1	0	1	0	1	0
B	0	1	0	1	0	1	0	1	0	1

Figure 9.3: Vectorised Form of Lists A and B

Each slot in the vector represents the presence or absence of a particular item in the original list, hence vector A has a '1' in the odd-numbered slots and a '0' in the even-numbered ones, since the list on which it is based contains only odd numbers. In the case of the lists of manually-selected concordance lines, therefore, each vector would be 200 items long, as there were 200 lines from which to choose, and it would contain up to 20 '1' entries, with the remainder being '0'.

In order to test the dissimilarity of the manual analyses, the respondents' lists of concordance lines were vectorised as described above and then compared pair-wise using a standard dissimilarity measure which is based upon determining the angle between two vectors. The maximum dissimilarity would be that the two vectors are at right-angles to each other. When this angle (90°) is converted to its sine, the result is 1.0, the maximum possible dissimilarity score. Less dissimilar vectors would lie at smaller angles to each other and would receive a lower score, since the sine of the angle would be smaller. The vectorisation algorithm is described in detail in Kaufman & Rousseeuw (1990).

The results of the pair-wise comparison of the vectors were put into a diagonal matrix, such that the comparison of vectors 2 and 1 was presented first, then of vectors 3 and 1, then 3 and 2 and so on in a similar manner to the method used for the simple measure of common items in the matrices shown earlier. The set of vectors derived from the lines

chosen as representative produces the following matrix:

```
 2  0.989
 3  0.981 0.981
 4  0.993 0.985 0.989
 5  0.979 0.988 0.991 0.998
 6  0.999 0.968 0.946 0.985 0.995
 7  0.985 0.973 0.975 0.964 0.910 0.938
 8  1.000 0.954 0.998 0.993 0.999 0.893 0.973
 9  1.000 0.915 0.985 0.972 0.990 0.915 0.921 0.941
10  0.999 0.951 0.980 0.984 0.978 0.858 0.881 0.912 0.875
11  0.999 0.995 0.981 0.993 0.988 0.989 0.973 0.954 0.979 0.979
12  1.000 0.917 0.992 0.938 0.979 0.937 0.957 0.937 0.882 0.933 0.989
      1     2     3     4     5     6     7     8     9     10    11
```

Figure 9.4: Dissimilarity Matrix for Representative Lines

The lowest value found in this matrix is 0.858 (row 10 column 6), which is still indicative of strong dissimilarity. This suggests that there was very little agreement among the respondents as to which lines were representative. This impression is reinforced by the presence of entries with maximal dissimilarity (1.000), indicating that there was no agreement whatsoever between the two sets of lines selected, although these do all occur in column '1', that is, in comparisons involving the first respondent's selections, indicating that this respondent's concept of representative lines differed entirely from many of the others.

A similar matrix was created for the usable lines:

```
2  0.908
3  1.000 0.991
4  0.984 0.980 0.991
5  0.978 0.985 0.988 0.957
6  0.983 0.951 0.978 0.951 0.972
7  0.984 0.954 0.963 0.954 0.915 0.951
8  1.000 0.946 0.996 0.966 0.975 0.980 0.966
9  0.959 0.880 0.994 0.971 0.983 0.946 0.971 0.952
10 0.984 0.954 0.979 0.917 0.973 0.887 0.954 0.981 0.919
11 0.983 0.951 0.978 0.979 0.972 0.949 0.933 0.918 0.946 0.951
12 1.000 0.992 0.996 0.983 0.977 0.982 0.969 0.987 0.981 0.969 0.992
     1     2     3     4     5     6     7     8     9     10    11
```

Figure 9.5: Dissimilarity Matrix for Usable Lines

In this matrix, the smallest value is 0.88 (row 9 column 2), which again indicates a large degree of dissimilarity. As above, three cells indicate maximal dissimilarity and, again, all refer to the same respondent.

As a final test, the process described above was carried out for twelve sets of completely random numbers in the range of 1 to 200. The resulting matrix is shown below:

```
2  0.999
3  0.968 0.995
4  0.989 0.989 0.980
5  0.980 0.995 0.989 0.980
6  0.995 0.954 0.999 0.980 0.995
7  0.989 0.968 0.995 0.999 0.999 0.980
8  0.995 0.989 0.989 0.989 0.968 0.995 0.989
9  0.999 0.980 0.999 0.980 0.989 0.999 1.000 0.989
10 0.995 0.989 0.968 0.995 0.980 0.989 0.980 0.968 0.999
11 0.989 0.980 0.968 0.980 0.995 0.995 0.989 1.000 0.968 0.989
12 0.980 0.995 0.968 0.995 0.989 0.989 0.989 0.995 1.000 0.995 0.995
     1     2     3     4     5     6     7     8     9     10    11
```

Figure 9.6: Dissimilarity Matrix for Random Numbers

A straightforward visual comparison of the random matrix with the representative or usable matrix will indicate that there is very little difference in the values. This can be simply confirmed by looking at the totals for all the cells in each matrix:

Matrix	Total	Mean
representative	63.67	0.964697
usable	63.739	0.965742
random	65.172	0.987455

Table 9.7: Summary of Matrix Contents

9.7.2. Additional Correlation Testing

The lack of correlation between the manual analyses, demonstrated in the previous section, makes it difficult to identify the set of software parameters which provides the closest correlation to either manual analysis. Two possible solutions to this problem present themselves; one means would be employ an alternative method of evaluating the automatic analyses, while the other solution would be to reconfigure the correlation test so that it is capable of producing valid, informative results. In the next two sections, both these avenues will be explored.

9.8. Alternative Evaluation Methods

If we accept the proposition, put forward in the section 9.6, that representativeness is largely a surface phenomenon, then it should be possible to analyse automatically various aspects of a particular concordance line and so arrive at an objective measure of its representativeness. Given the currently available computational tools, two means of investigation spring to mind. The first would involve determining the proportion of citations which contained no significant collocates of the node word in question. The second would measure the percentage of lines with unusual syntactic structure.

While no testing has been undertaken using syntactically analysed data, the *cohort* software relies heavily upon the presence of repeated contextual features in order to produce its results. This is analogous to collocational analysis, although some configurations use a stricter definition of collocation than is generally accepted, in that they demand that the collocates be in a specific position relative to the node word. If the loosest definition of

contextual feature is used (collocates may appear anywhere within a fixed context of the node word), this will provide the closest parallel to conventional collocational analysis. The proportion of lines which we might expect to become linked can then be established on the basis of the density of significant collocates in concordance lines. To test this hypothesis, an analysis of this kind was carried out on the full concordance (16,818 lines) of 'date'.

9.8.1. Identifying Collocates

For each word in the context provided by the concordance line, a Z score was calculated on the basis of the ratio between word's observed and expected† co-occurrence with the node word 'date'. The significance level was defined as an observed:expected co-occurrence ratio of 2.0 or greater; that is, the collocates had to occur with the node at least twice as frequently as would be expected by chance. Using this minimum Z score of 2.0, it was noted that 16,803 lines (99.91%) contained at least one significant collocate. This corresponds very closely to the results obtained using a link threshold of one and the raw link type (199/200 or 99.5%), suggesting that the similarity proposed earlier between traditional collocates and raw links is a close one. The results of the collocational analysis can be found in the table which follows:

---

† The expected frequency of co-occurrence was calculated on the basis of the overall frequency of each word in the BoE, as extracted from a wordlist supplied by Jeremy Clear of Cobuild. It is equivalent to

frequency of collocate in corpus × context size + corpus size

where the size of the context and the corpus is defined in terms of the number of running words (tokens) that they contain.

Significant Collocates	Number of lines	Per Cent
0	15	0.10
1	66	0.39
2	190	1.13
3	514	3.06
4	962	5.72
5	1400	8.32
6	1862	11.07
7	2051	12.20
8	2055	12.22
9	1986	11.81
10	1553	9.23
11	1285	7.64
12	974	5.79
13	720	4.28
14	565	3.36
15	293	1.74
16	173	1.03
17	75	0.45
18	59	0.35
19	16	0.10
20	2	0.01
22	2	0.01
Total	16,818	100.01

Table 9.8
Count of Significant Collocates in 16,818 Complete Concordance lines

Compare this now with the results using just the 200 lines which were presented to the respondents.

Significant Collocates	Number of lines	Per Cent
1	2	1
2	4	2
3	9	4.5
4	18	9
5	23	11.5
6	27	13.5
7	22	11
8	29	14.5
9	12	6
10	22	11
11	12	6
12	8	4
13	6	3
14	2	1
15	4	2

Table 9.9
Count of Significant Collocates in 200 Complete Concordance lines

To see how the two sets of collocates compared, the percentages from the above two tables were plotted on a graph:

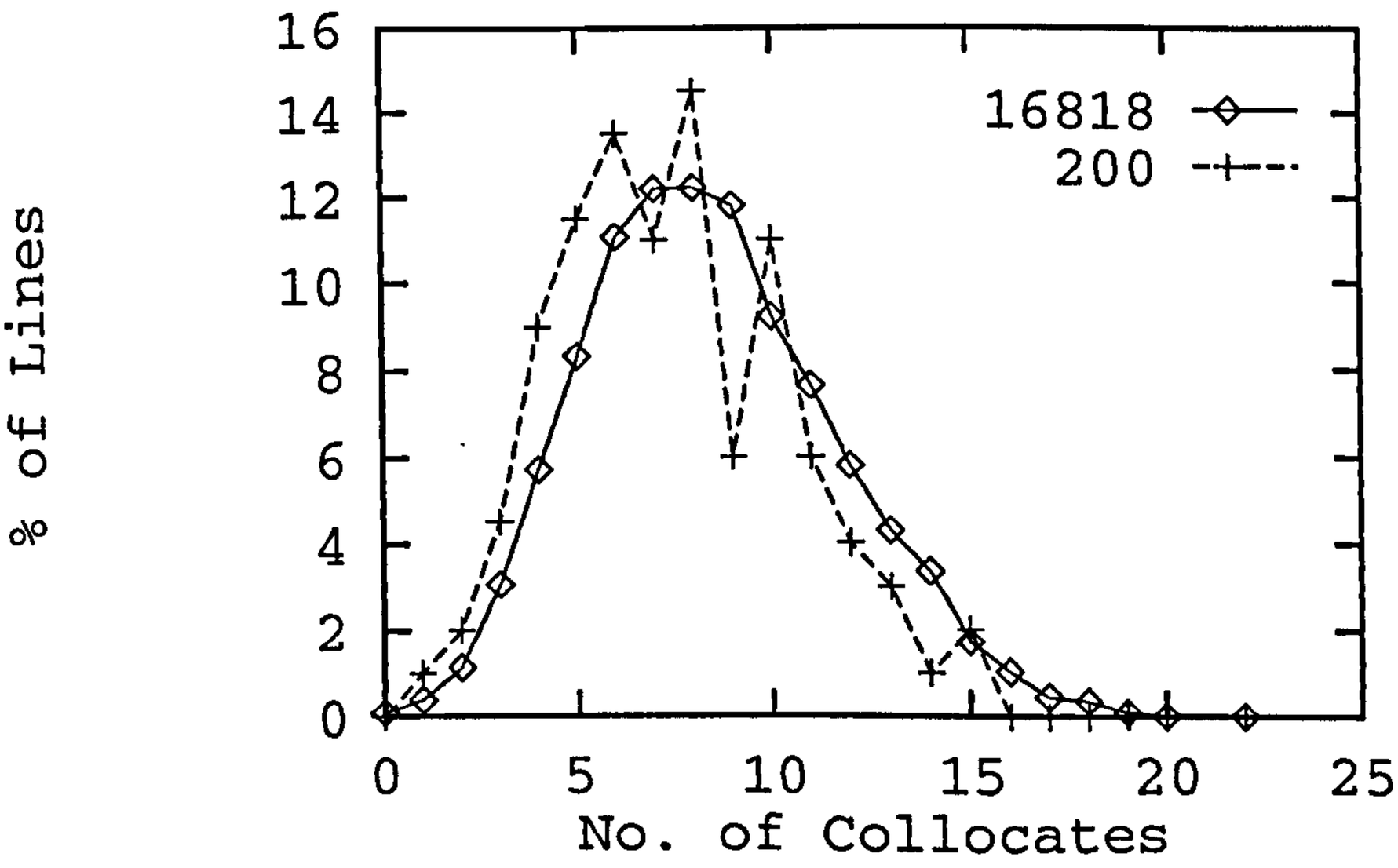

Figure 9.7
Comparison of Collocate Counts: 16,818 lines vs 200 lines

On the whole, the two plots follow a similar line, with the "16,818" plot tailing off slightly later and the "200" plot showing slightly greater variation, especially around the higher values. The greater variation on the "200" plot is no doubt due to the size of the samples involved, since in the 200-line sample a change of only one line is enough to produce a 0.5% variation, whereas a change of this magnitude would represent over 80 lines in the full 16,818-line concordance. The fact that the "16,818" plot tails off more gradually is explained by the low probability of the occurrence of lines containing more than 15 significant collocates – this set is 80 times the size of the 200-line sample and the likelihood of it containing lines with this many significant collocates is correspondingly greater.

For the purposes of comparison, the same exercise was carried out using a collocational context of ±4 words. In the table below we see a summary of the number of significant collocates that occur in each line in the complete concordance of 'date', drawn from the BoE. The table indicates that there are 2,659 lines out of the total 16,818 which contain no significant collocates at all, 456 which contain only one significant collocate and so on. With a restricted span being used in identifying the collocates, we can obtain a better picture of the density of collocation for a given line, i.e. what proportion of the words in its environment are significant collocates.

Significant Collocates	Number of lines	Per Cent
0	2,659	15.81
1	456	2.71
2	1,507	8.96
3	2,893	17.20
4	3,412	20.29
5	2,912	17.31
6	1,859	11.05
7	820	4.88
8	300	1.78
-	16,818	100

Table 9.10: Count of Significant ±4 Collocates in Concordance lines

9.8.2. Collocates in the Automatic Analyses

Using the list of significant collocates established for the unrestricted span in the previous section, a score was assigned to each line in the 'date' sample on the basis of the number of significant collocates which it contained. Across the sample as a whole, a total of 1,494 significant collocates were found, giving a mean significant collocate count of 7.47 across the 200 lines. The maximum score for any one line was 15 and the minimum 1.

The scores obtained in this manner were applied to the lines selected by the respondents. For the 82 lines rated as usable, 661 significant collocates were counted, representing an average of 8.06 significant collocates per line and for the 92 representative lines 741 were found, giving a mean of 8.05. The fact that both these averages are higher than the overall mean for the full 200 lines would seem to indicate that the presence of significant collocates is playing a part in the manual selection of concordance lines.

The same comparison was then made for each set of automatically-selected lines; a sample of the results are shown in the table which follows.

No. of Links	Link Type	Number of Collocates	No. of Lines	Max Collocates	Min Collocates	Average Collocates
1	raw	1,488	199	15	1	7.48
1	abs	1,424	188	15	1	7.57
1	rel	1,477	197	15	1	7.50
2	raw	1,206	150	15	1	8.04
2	abs	353	38	15	3	9.29
2	rel	1,205	150	15	1	8.03
3	raw	440	46	15	2	9.57
3	abs	97	9	15	5	10.78
3	rel	388	41	15	3	9.46

Table 9.11: Significant Collocates for Automatic Analyses

In all cases, the automatically-selected lines achieve an average collocate count which is higher than the overall mean (7.47). Comparing these figures with the results from the manual selections, the closest match to the average for the lines selected by the informants occurs when a link threshold of 2 is used with the raw or relative link type.

This data gives a clear indication that lines containing a higher number of links have a higher average number of significant collocates. The link type is also a factor in this relationship, as for each link threshold value it is always the absolute link type which attracts the highest average collocate count, while there is little to distinguish the raw and relative figures.

The evidence presented in this section would seem to support the hypothetical link between cohesion and representativeness, since it has been shown that collocation plays a part in the manual selection of lines and also that there is a correlation between the presence of significant collocates and certain software parameters. This chain of reasoning reinforces the possibility of a connection between the manual selections and the set of parameters supplied to the automatic system, a connection which we attempted to identify in Section 9.7, but failed to do because of the lack of a suitable correlation measure. In the section which follows, a range of measures will be introduced which aim to address this lack.

9.9. Re-evaluation of the Sample

The fact that the automatic system was used to analyse only 200 lines is not a fair means of assessing its usefulness, since the intention is to run the software over amounts of corpus evidence that are in fact too large for a human being to be able to manipulate. A fairer test would therefore be to analyse all the available concordance lines for 'date' and then examine the scores for the 200 lines included in the sample *on the basis of evidence from all the lines*. This is, after all, closer to the task which the corpus users perform, since they will not have evaluated the 200 lines solely on the evidence presented by the sample set, but rather in accordance with external knowledge which they have of the node word's behaviour. This is evidenced by the fact that several respondents suggested that there were senses missing from the sample.

The application of this approach necessitated the automatic analysis of the original full 'date' concordance. Before the analysis could proceed, these lines had to be 'sanitised', which mainly involved the removal of duplicate or near-duplicate lines (e.g. those which differed only in their punctuation), but also required the re-insertion of some of the sample lines, where these had not been included in the original. † The sanitised concordance set consisted of 16,761 lines and the 200 lines which made up the sample were placed at the start, so that their line numbers in the full set and the sample would correspond. The concordance was then processed using the range of parameters described previously, resulting in the customary 252 sets of output. In the following sections, several methods for evaluating these sets of output relative to the manual analyses will be introduced.

9.9.1. Simple Ranking

As we saw at the end of Section 9.7, there is no statistically identifiable correlation between the various sets of manually selected lines, yet it has already been noted, in sections (9.4.1.3 and 9.4.2.3), that there is a visible degree of overlap between the respondents' selections in the case of the most popular lines. To resolve this paradox, a less complicated approach was taken to the evaluation of the new sets of automatically-selected lines, one which made use of the simple questionnaire data presented in condensed form in the 'Results in Detail' sections (9.4.1.2 and 9.4.2.2).

In Tables 9.1 and 9.2 the lines selected manually were presented in descending order of the number of votes they received from the respondents. In order to evaluate the success of the automatic selection system, a measure was devised which applied the ranking of the top-scoring items from each of the tables to the various automatic analyses. If the lines chosen by several respondents appeared near the top of a set of automatic output, this would indicate that the set of parameters used to generate it was successfully identifying lines which would be preferred by a corpus user. This measure also incorporates the

† Both the sample set and the full concordance were generated directly from the Cobuild corpus database, i.e. the sample was not derived from the full set. The reason for the absence of some of the sample lines from the full set is possibly attributable to Cobuild's policy of removing older material from their corpus.

position of the manually-selected lines within the automatic analyses; since the concordance lines presented by the software are ranked in descending order of the bonds that they have acquired, a stronger match between automatic and manual analyses will be achieved when more top-ranking manually-selected lines occur nearer to the top of the automatic output. The strongest match would therefore occur when all the manually-selected items occurred at the very top of the automatically-ranked set of lines.

The means by which the match between manual and automatic analyses was calculated is as follows. For each set of automatically-selected lines, a raw score was calculated based on the rank of the highest-scoring manually-selected items within the set. Supposing that the top ten manual lines were being used, then the lowest possible score would be achieved if these ten lines appeared as the first ten lines of the automatic output, representing the best possible match and resulting in a raw score of 55 (1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9 + 10). In cases where a line chosen manually is not selected by the software, a score of 201 is given (because there are 200 lines), making the maximum raw score 2,010, which would indicate that none of the ten lines were selected by the software. Since we wish to have a measure of the strength of similarity between the two analyses, where the best match attracts the highest score, the raw scores were rescaled and inverted so that a score of 1 would indicate a perfect match and a score of zero would be indicative of the poorest match possible. This new, 'simple ranking' score was calculated according to the formula:

$$score = 1 - \frac{x - \min}{\max - \min}$$

where x represents the raw score, *max* the maximum possible raw score and *min* the minimum possible raw score.

9.9.1.1. Comparison with Representative Lines†

Let us start by reviewing the lines selected by the respondents as being most representative. The following table gives the manual scores received by the top ten lines, which all received five or more votes:

Line No.	No. of Votes
159	9
57	7
6	6
55	6
4	6
143	6
87	5
65	5
25	5
174	5

and the lines themselves are shown below, in the same order:

```
cers so they can keep their members up to date with what is happening in the industry.2
 periods of deep loneliness and grief.Her date of birth has been placed somewhere aroun
s had declared their willingness to set a date for starting stage two of economic and m
dents had slept with a man on their first date and 39 per cent admitted to being unfait
ontinue to try to get you into bed. <LTH> Date rape is at the forefront of all our mind
jor is expected to confirm April 9 as the date of the election. <h> Tories pin election
r information was always six hours out of date. I get an update from the senior forecas
tempt the same operation again at a later date. He may even, some analysts say,risk the
ese words and so we will see that sell-by date is no longer associated with perishable
In India where we've got reasonably up to date statistics on population. Er there's bee
```

In the three tables which follow, a sample of the simple ranking scores for the various sets of output is presented. The tables are drawn from the top, middle and bottom of the total results, sorted by descending score, hence the parameter combinations which achieved the best match will be found in the first table. The other tables are provided to give an impression of the range of scores and the full listing of all the parameter combinations is located in Appendix 5b †.

Comparison with usable lines has been omitted.

† A similar table based on a comparison with lines chosen as usable can be found in Appendix 5c.

Stopword List	No. of Links	Span Type	Link Type	Score
arts-prons	2	open	abs	0.786701
arts-prons	4	open	raw	0.781586
arts-prons	3	open	raw	0.781074
arts-prons	2	open	raw	0.776471
arts-prons	3	open	abs	0.765729
arts-prons	2	open	rel	0.764706
arts-prons	3	open	rel	0.760614
top50	1	open	abs	0.757033
arts-prons	1	open	raw	0.755499
bt	1	open	abs	0.748849

Table 9.12a
Simple Rank Scores based on Top 10 Representative Lines
Top Ten Items

Stopword List	No. of Links	Span Type	Link Type	Score
btb	3	fixed	raw	0.313555
zero	4	fixed	abs	0.308951
arts-prons	4	fixed	rel	0.299744
top100	2	fixed	raw	0.287468
top100	2	fixed	rel	0.287468
zero	6	open	abs	0.286957
arts-prons	5	open	abs	0.280818
top100	2	fixed	abs	0.273657
bt	2	fixed	rel	0.2711
top150	3	open	rel	0.258824

Table 9.12b
Simple Rank Scores based on Top 10 Representative Lines
Middle Ten Items

Stopword List	No. of Links	Span Type	Link Type	Score
top150	6	fixed	rel	0
top50	4	fixed	abs	0
top50	4	fixed	raw	0
top50	4	fixed	rel	0
top50	5	fixed	abs	0
top50	5	fixed	raw	0
top50	5	fixed	rel	0
top50	6	fixed	abs	0
top50	6	fixed	raw	0
top50	6	fixed	rel	0

Table 9.12c
Simple Rank Scores based on Top 10 Representative Lines
Bottom Ten Items

A cursory glance at the highest-scoring items in the list suggests that the 'arts-prons' stopword list performs well, especially in combination with the open span, relative or raw links and a mid-range link threshold. This can be confirmed by calculating the proportion of the total score achieved by each possible value of each parameter. The results from this are shown in the table below, which indicates, for example, that parameter combinations which include the 'arts-prons' stopword list make up 21.6% of the total score (85.6952) and therefore account for the largest proportion of the total, closely followed by the 'zero' stopword list at 21.5%.

Parameters							
Stopword List		Link Threshold		Span		Link Type	
bt	10.6	1	31.0	open	63.2	Abs	29.1
btb	13.8	2	25.4	fixed	36.8	Rel	34.2
arts-prons	21.6	3	18.6			Raw	36.7
top150	9.5	4	11.7				
top100	10.2	5	7.6				
top50	12.8	6	5.7				
zero	21.5						

Table 9.13: Summary of Scores: Representative Lines

In the link threshold column it can be seen that it is actually '1' which dominates, although '2' and '3' are also strong and occupy most of the top seven positions in Table 9.14a.

Let us now move on to look at the output for the highest scoring parameter combination: 'arts-prons' stopword list, 2 absolute bonds and an open span, or 'arts-prons/2/open/abs', to use a convenient shorthand. The next table repeats the ranked listing of the top ten representative lines, with the addition of their raw position, rank and bond score from the automatic analysis. Since some lines acquired the same scores, the 'Position' column does not give a complete picture of the relative rank. To circumvent this problem, the 'Rank' scores (manual and automatic) are based on putting all lines with the same score into one 'bucket'. In the case of the automatically-derived ranks, bucket 1 lines score 1688 bonds and bucket 149 lines score no bonds (although it so happens that both the first

and last buckets contain only one line).

It can be seen that all of the manually-selected lines are also selected by the software when using this parameter combination, although the automatically-derived ranking bears little resemblance to the manual ranking by which the items have been ordered.

Line No.	Manual Rank	Auto. Rank	No. of Bonds	Position
159	1	7	1550	7
57	2	42	276	44
6	3	40	302	42
55	3	65	174	68
4	3	83	116	90
143	3	76	140	80
87	4	25	590	25
65	4	36	325	38
25	4	63	179	66
174	4	12	1483	12

Table 9.14
Comparative Ranks: Representative vs arts-prons/2/open/abs

The ten lines which scored most highly using parameter combination arts-prons/2/open/abs, shown here with their line numbers and bond score, were:

```
168 1688 to be different. <LTH> As we come up to date, people # do it'' to be the same. L
177 1651 lining my 'firm grasp of the most up-to date trauma procedures". <t> The referen
156 1629 ve to enable the returner to keep up to date with developments.<t> Various other
157 1617 ought of her man # <SO> Very much up to date, only been in service with our own
154 1615 nd he said he would bring Mr Bush up to date on the issue:If we were forced to r
152 1581 retary of state, brought Franklin up to date on the bloodshed in his beloved Fra
159 1550 rs so they can keep their members up to date with what is happening in the indus
188 1548 rld, and though this is the first up-to date survey of its politics, it does not
153 1512 ut adequate nuclear weapons, kept up to date and based forward in Europe, our de
173 1495  really because er it's just been up-to date and it <M01> Mhm.<M02> I mean that'
```

The most obvious characteristic of these lines is that they all exemplify the same phrase, 'up to date' (where this occurs with the optional hyphen, the software has split it into two words). The bond scores for these lines suggest that this is a very frequent phrase and this is confirmed by the concordance data, which contains 738 occurrences of 'up to/up-to date', 15 of which occupy the top positions in the ranking. Furthermore, the wordlist for this concordance tells us that 'up' occurs 1,419 times in the -2 position and that 'to' is to be found in the -1 slot 4,622 times.

The obvious pattern in the next most high-scoring positions in the output from arts-prons/2/open/abs is the phrase 'out of/out-of date', occupying positions 16-25 in the ranking and occurring 461 times in the total concordance. All the lines from the sample are presented in descending order of bond score in Appendix 6.

More generally, the parameter combination delivers a fine-grained ranking, with 149 discrete bond scores ranging from the values seen above right down to zero, which is applied to just one line.

9.9.2. Pearson Correlation 1

By using only the top few manually-selected lines, which were chosen by more than one respondent, it was hoped that the problem of the lack of correlation among the manual sets would be circumvented. One drawback of this approach, however, is that it does not take account of the relative ranking of the top-scoring lines.

As noted earlier, the Pearson Correlation Coefficient test can be used to measure the degree of correlation between the relative ranking of items on two lists. The correlation tests were therefore repeated using only the subset of items just described. This gave rise to some highly significant† results, which are presented below.

Stopword List	No. of Links	Span	Link Type	r
arts-prons	6	fixed	abs	0.845154
arts-prons	6	fixed	raw	0.845154
arts-prons	6	fixed	rel	0.845154
arts-prons	4	open	rel	0.82091
arts-prons	4	open	abs	0.801658
arts-prons	3	fixed	abs	0.783591
arts-prons	3	open	rel	0.75107
arts-prons	4	fixed	abs	0.750098
arts-prons	3	open	abs	0.749034
arts-prons	5	fixed	abs	0.742307
arts-prons	5	open	rel	0.731502
arts-prons	5	open	abs	0.724418

† For this version of the Pearson's test, with N=10, the smallest correlation coefficient which would be significantly different from zero at the 0.05 level is 0.55.

Stopword List	No. of Links	Span	Link Type	r
arts-prons	4	fixed	raw	0.703508
arts-prons	6	open	abs	0.698923
top50	2	open	abs	0.686517
arts-prons	3	fixed	raw	0.6744
zero	5	open	abs	0.661859
zero	5	fixed	abs	0.651644
arts-prons	5	fixed	rel	0.633866
zero	6	open	abs	0.617854
top50	2	open	rel	0.616413
arts-prons	1	open	abs	0.608489
arts-prons	3	fixed	rel	0.602194
arts-prons	4	fixed	rel	0.592667
arts-prons	5	fixed	raw	0.582955
arts-prons	6	open	raw	0.578729
arts-prons	5	open	raw	0.576099
arts-prons	1	fixed	abs	0.56551

Table 9.15
Pearson Scores for Top 10 Representative Lines vs Automatic

On the basis of this data, the parameter combinations which contain the 'arts-prons' stop-word list appeared to perform best, taking part in 19 out of the 28 correlating analyses. Tempting as it is to accept these results at face value, the fact that lines acquiring six links within a fixed span achieved the closest correlation ought to be a source of concern, since the implications that this would have for the supposed behaviour of the node word in question do not correspond at all with linguistic intuition, simply because of the implausibility of these lines forming links via six out of the possible eight words in the node's context.

Upon closer examination it became clear that, with only one exception, none of the 'top' manually selected ten lines actually acquired any bonds; that exception was the top-scoring line (159), which acquired two bonds, as can be seen in the following figure, which gives the numnber of bonds generated using the arts-prons/abs/6/span parameters:

Line Number	Manual Score	Automatic Score
159	9	2
57	7	0
4	6	0
6	6	0
55	6	0
143	6	0
25	5	0
65	5	0
87	5	0
174	5	0

This phenomenon has a profound effect on the calculation of correlation, since the method used relies on comparing the difference in rank across the two sets of data. Where there is virtually no variation in the ranking within one of the sets, the method fails to function, although there is no way of detecting this, other than checking the raw figures, before the correlation is calculated. Several other parameter sets fell victim to this problem and unfortunately, the presence of these zero values invalidates this type of correlation test, since all items must have a non-zero score in both the sets under comparison. Those parameter combinations which were not found to be invalid for this reason are shown in the next table:

Stopword List	No. of Links	Span	Link Type	r
arts-prons	4	open	rel	0.82091
arts-prons	3	open	rel	0.75107
arts-prons	3	open	abs	0.749034
arts-prons	3	fixed	raw	0.6744
top50	2	open	rel	0.616413
arts-prons	1	open	abs	0.608489
arts-prons	5	open	raw	0.576099
arts-prons	1	fixed	abs	0.56551

Table 9.16
Validated Pearson Scores for Top 10 Representative Lines vs Automatic

These results correspond more closely to what one might expect. The 'arts-prons' stopwords still predominate, but the other parameters appear in less strict combinations than was seen in the previous table. The restricted span appears only twice and combines either with the loosest link type and three links, or with the absolute link type but only

one link.

Examination of the raw scores for the most strongly correlating pair of analyses (manual vs arts-prons/4/open/rel) reveals a correlation which is apparent even without statistical analysis:

Line Number	Manual Score	Automatic Score
4	6	55
6	6	148
25	5	89
55	6	70
57	7	76
65	5	52
87	5	31
143	6	82
159	9	270
174	5	82

The ten top-scoring lines using arts-prons/4/open/rel are given below:

```
 75 484  ministry denies there is a hold-up, no date has been set for a new round of talks
163 479 o teach has not been very productive to date, nor is it likely to become more so i
 79 472 not letting us go to work.<t> Dugan: No date has been set for the resumption of co
161 444 oy In New Cross # their greatest hit to date, is nowhere to be seen; but they do s
157 424 ought of her man # <SO> Very much up to date, only been in service with our own fo
168 410 to be different. <LTH> As we come up to date, people # do it'' to be the same. Lon
  7 6 408 n population is a minority in Serbia.No date has been set for elections and there
175 387 vities in which you've been involved to date?<M01> Er the spectrum of that would r
188 301 rld, and though this is the first up-to date survey of its politics, it does not p
191 282 unt of ownership.<t> Levinson: No trial date has been set yet for the Janis lawsui
```

As might be expected from the use of the relative link type, the patterns represented here are more diverse than those seen for the arts-prons/2/open/abs combination which we examined earlier. In addition to the 'up to/up-to date' pattern, there is also 'no date has been set', as well as 'to date' in the 'so far/until now' sense. Line 75 may well be the product of over-rich bonding, since it incorporates the pattern 'date has been set', but will also have acquired bonds on the basis of the 'up' in slot -2, which will have linked to lines with the 'up to date' pattern.

As far as the range of the bond scores achieved by this parameter combination is concerned, 100 different scores were applied across the 200 line sample, which is an even coarser ranking than we saw for arts-prons/2/fixed/abs. Additionally, 19 lines received no bonds at all.

9.9.3. Pearson Correlation 2

Having established that the Pearson test was capable of producing results which were statistically significant and corresponded to linguistic intuition, the test was extended to include all the items from the 200-line sample which were chosen by more than one respondent.

54 lines remained once the singleton items had been removed. Using this set of lines as the basis for comparison, nine of the automatic analyses correlated significantly with the manual selection, scoring more than the threshold of 0.271. These are given in the next table. As before, the statistically invalid combinations have been included in the lower half of the table, as these still have some linguistic interest.

Stopword List	No. of Links	Span	Link Type	r	Valid
arts-prons	3	open	raw	0.487403	y
arts-prons	2	open	raw	0.432735	y
arts-prons	2	open	rel	0.358678	y
arts-prons	1	open	rel	0.338473	y
arts-prons	3	open	rel	0.327846	y
arts-prons	1	open	raw	0.324546	y
arts-prons	4	open	raw	0.447633	n
arts-prons	3	fixed	raw	0.33147	n
arts-prons	3	fixed	abs	0.328121	n
arts-prons	3	open	abs	0.320494	n
arts-prons	5	open	raw	0.273281	n

Table 9.17: Pearson Scores for Representative Lines vs Automatic

In this set of data, the 'arts-prons' stopword list is the only one present and combines exclusively (for the valid items) with the open span and with link thresholds and types that are far from strict.

Since the number of items being correlated is somewhat higher, the raw figures for manual and automatic scores are not given here, as no correlation would be manually identifiable. We shall instead move on to the lines which acquired most bonds with arts-prons/3/open/raw:

```
157 5306 ht of her man # <SO> Very much up to date, only been in service with our own
163 4708 each has not been very productive to date, nor is it likely to become more so
175 4703 ies in which you've been involved to date?<M01> Er the spectrum of that would
 93 4187 e next three weeks # The NBL cut-off date for the finalisation of imports is
 79 3947  letting us go to work.<t> Dugan: No date has been set for the resumption of
 75 3806 nistry denies there is a hold-up, no date has been set for a new round of tal
152 3776 ary of state, brought Franklin up to date on the bloodshed in his beloved Fra
 58 3679 that it is vital.<p> Dr Salk's ideas date back a long way, but he has linked
140 3565  the--on the--petrified tree and the date # And in all of my trips out to Mon
189 3045 kes it the No. 1 film of the year to date and the biggest April release in th
```

In these lines, the '(up) to date' pattern is still present to some extent, but is diluted by other patterns, introduced by the use of the raw link type. The overall level of bonding is still high, as evidenced by the values shown in these few lines and confirmed by the fact that there are 193 different bond scores in the analysis, with the lowest score being 2.

9.9.3.1. Usable Lines

Of the lines selected as usable, 49 were chosen by more than one respondent. Only three of the automatic analyses were found to correlate significantly ($r >= 0.284$) with this subset of lines; these analyses are listed in the table below, with an indication of their validity according to the criterion mentioned earlier.

Stopword List	No. of Links	Span	Link Type	r	Valid
arts-prons	2	fixed	rel	0.306408	y
arts-prons	2	open	abs	0.29841	y
arts-prons	2	fixed	abs	0.306152	n

Table 9.18: Pearson Scores for Usable Lines vs Automatic

Once again, it is the 'arts-prons' list which provides the best correlation to the manual results. In the two valid combinations shown here, it combines with other parameters which represent a compromise of span and link type, while the link threshold remains

constant.

The top ten lines using the arts-prons/2/fixed/rel combination are:

```
188 1887 ld, and though this is the first up-to date survey of its politics, it does not
157 1853 ught of her man # <SO> Very much up to date, only been in service with our own
159 1840 s so they can keep their members up to date with what is happening in the indus
176 1831 e now available but ask somebody up-to date.<M01> Mm.<F01> And of course comput
152 1761 etary of state, brought Franklin up to date on the bloodshed in his beloved Fra
168 1703 o be different. <LTH> As we come up to date, people # do it'' to be the same. L
184 1677 en though the home loan was paid up to date. Few would ever have imagined they
153 1632 t adequate nuclear weapons, kept up to date and based forward in Europe, our de
177 1598 ining my 'firm grasp of the most up-to date trauma procedures". <t> The referen
154 1595 d he said he would bring Mr Bush up to date on the issue:If we were forced to r
```

The lines shown here bear a strong resemblance to the set shown earlier for the arts-prons/2/open/abs combination, with eight out of ten items in common. This parameter set, arts-prons/2/open/abs, is coincidentally the second most strongly correlated combination for this comparison. In both the sets in this test, the dominating pattern is 'up to/up-to date', which occupies the first 15 positions in arts-prons/2/fixed/rel. Strong patterns further down the ranking include 'out of date', 'date of birth', 'to date' and 'date ... set'.

A good range of bond scores are attached to the lines, with 154 discrete values and only four zero-bonded lines.

9.10. Conclusions

In the course of re-evaluating the concordance line sample, we performed three correlation tests on the two sets of manually-selected lines. With a view to summarising the effects of the various parameter combinations, let us examine again the combinations which achieved the best match for each test and data set. In the table which follows, 'Repr' refers to the concordance lines chosen as representative (or some subset thereof as defined earlier) and 'Use' denotes the lines selected as usable.

Test	Compared with	Parameter Set			
		Stopwords	Links	Span	Link Type
Ranking	Repr	arts-prons	2	open	abs
Ranking	Use	arts-prons	2	open	abs
Pearson 1	Repr	arts-prons	4	open	rel
Pearson 1	Use	zero	1	fixed	abs
Pearson 2	Repr	arts-prons	3	open	raw
Pearson 2	Use	arts-prons	2	fixed	rel

Table 9.19: Summary of Best-Match Parameter Combinations

On the evidence of this summary, it seems safe to conclude that in general the 'arts-prons' stopword list produces the best correlation between the manual and automatic analyses. To some extent, this echoes the Summary of Scores tables presented in the sections on simple ranking comparison, where 'arts-prons' accounted for the highest proportion of the scores, closely followed by 'zero', which is also represented here, albeit only once.

The open span is generally superior, but the fixed span seems slightly better when usability is the issue. The ratio of open to fixed shown here, 2:1, is very similar to that seen in the Summary of Scores tables, where the fixed span performed marginally better for usability than for representativeness.

The remaining parameters seem to combine to create a compromise in terms of strictness, so that if the absolute link type is employed then a low link threshold is used, for example. Similarly, high link thresholds are avoided where the fixed span is used and when the fixed span is combined with the absolute link type, the lowest possible link threshold is used.

It is interesting and encouraging to note that these combinations of parameters display a high degree of overlap with the ten identified as candidates at the end of Chapter 7 and presented in Chapter 8, more so because these related to a different set of concordance lines (for the node word 'exchange'). All but one of the 'exchange' combinations made use of either the 'arts-prons' or 'zero' stopword lists and seven of the ten used the open span. In addition, a number of the top 'exchange' combinations were present amongst the

highest scoring items in the simple rank and Pearson correlation tests for 'date'. These are detailed in the Table 9.20.

cccc. Parameter Combination	Comparison	Table Reference	Position in Table
arts-prons/1/open/raw	Simple Rank vs Representative	9.12a	9
zero/1/fixed/raw	Pearson 1 vs Usable	9.20	2
zero/1/fixed/rel	Pearson 1 vs Usable	9.20	3
arts-prons/1/open/rel	Pearson 2 vs Representative	9.22	4
arts-prons/1/open/raw	Pearson 2 vs Representative	9.22	6

Table 9.20: Results for high-scoring 'exchange' combinations

As a final summary of the relative performance of the various parameters, the next two tables show the number of times that each parameter value appears in each of the outputs from the Pearson tests as a valid, significant correlation. This is similar in principle to the Summary of Scores tables seen earlier, but this time simply records the frequency of occurrence of each parameter value, since adding Pearson scores across different sets of results is not a valid operation. The first table shows the results for the two Pearson tests which were used to correlate the automatic analyses with the lines manually-selected as representative.

Parameters							
Stopword List		Link Threshold		Span		Link Type	
bt	0	1	4	open	12	Abs	3
btb	0	2	3	fixed	2	Rel	6
arts-prons	13	3	5			Raw	5
top150	0	4	1				
top100	0	5	1				
top50	1	6	0				
zero	0						

Table 9.21
Occurrences of each parameter value in Pearson 1 & 2 – Representative

In the two Pearson tests for representativeness, a total of 14 parameter combinations were found to correlate with the manual analysis. The summary given above confirms the earlier suggestion that the 'arts-prons' list provides the closest match, occurring in all but one of the correlating analyses. The 'open' span, similarly, occurs in all but two of the

combinations and this also reinforces the conclusions drawn earlier on the basis of the Summary of Best-match data. With so little variation in their values, it is difficult to see a significant pattern in the link threshold figures, although values under four would appear to perform best. Similarly, the link type data suggests that the looser values are to be preferred over the absolute type.

Moving on now to look at the summary for the Pearson tests for correlation with the usable lines:

Parameters							
Stopword List		Link Threshold		Span		Link Type	
bt	0	1	7	open	4	Abs	5
btb	0	2	5	fixed	8	Rel	5
arts-prons	6	3	0			Raw	2
top150	0	4	0				
top100	0	5	0				
top50	0	6	0				
zero	6						

Table 9.22
Occurrences of each parameter value in Pearson 1 & 2 – Usable

This set of data presents quite a different picture of the most suitable parameters from the results for representativeness. The 'zero' stopword list is placed on an equal footing with the 'arts-prons' one and the fixed span moves into first place with twice the occurrences of the open span. The link threshold now shows a definite tendency towards the lower values and the preferred link types have shifted toward greater strictness.

The conclusions that can be drawn from this summary data go some considerable way towards determining the optimal initial settings for the identification of representative or usable lines and provide strong clues as to how the two categories might be distinguished. In particular, the high incidence of a minimal or empty stopword list and a fixed span in the parameter combinations which correlate best with the lines regarded as usable, suggests that the respondents, in selecting usable lines, were favouring paradigmatically and syntagmatically defined patterns which incorporated high frequency items and occurred at close range to the node word.

In addition, we have seen how altering the parameters can be used to isolate lines with particular features, such as increasing the link threshold and reducing the span if the corpus user's aim is to identify lines which exemplify fixed phrases.

Sadly, this may prove to be an unrepeatable experiment, since the editorial team which assisted me with the manual evaluations of the sample concordance has now been dissolved and it will be some time before such a high concentration of professional corpus users with so much experience is reconstituted.

Chapter 10

Future Research

10. Future Research

10.1. Summing Up

The motivation for this study grew out of the author's interest in two strands of linguistic research which, up until now, might have been regarded as quite distinct from each other. The first strand, corpus-based analysis, deals most commonly with substantial amounts of text and seeks to derive generalisations from the (increasingly) large body of evidence which the corpus presents; with regard to the storage, display and, to some extent, automatic analysis of its underlying data, the corpus-based approach is highly amenable to computerisation, indeed would never have achieved the scale and popularity which it now enjoys without the use of computers. The other strand, the study of lexical cohesion, operates at the level of the individual text and concerns itself with phenomena which are difficult or impossible to identify using computational means; only in the area of automatic abridgement has this analytical approach proved itself to be automatable, but even then it employs only a small subset of the cohesive features which would be recognised by a human analyst.

This thesis has attempted to interweave these strands with the aim of creating a means of dealing with the information overload which users of large corpora now face and which threatens to undermine the usefulness of continued corpus expansion. In Chapter 1 we saw that the keyword-in-context concordance is still a centrally important tool for corpus users, while the drawbacks involved in its use on very large corpora were introduced in Chapter 2. Chapter 3 examined the parallel features between a concordance and a conventional text and introduced the idea that the identification of cohesive ties between concordance lines and the subsequent creation of an 'abridgement' of them would be a reasonable and possibly productive line of research in which to engage. Chapter 4 outlined the basic methodology of lexical cohesion analysis and its applicability to automatic abridgement.

Chapter 5 covered the functionality of the concordance line selection software, *cohort*, and demonstrated its relationship to the original abridgement software, with a special focus on the additional parameters which had been included in its design in order to allow the software to be run on data derived from concordance lines. The two following chapters described firstly the individual parameters to the software and secondly the ways in which these interact, as well as suggesting that certain combinations of parameters might be more useful than others. A sample of the output produced by *cohort* using these candidate parameter combinations was then presented in Chapter 8.

Having established a system for selecting concordance lines, it is, of course, essential that it should be validated in some way. This issue was addressed in the rather substantial Chapter 9, in which the opinions of expert corpus analysts were employed to identify the optimal combination of parameters for identifying concordance lines which met either of the criteria of representativeness or usability. It is unfortunate that the difficulties involved in soliciting expert assistance in the manual evaluation of the concordance lines, coupled with the vastness of the results that are generated, even when varying just the handful of parameters which we have examined in the previous chapters, were such a limiting factor on the scope of the results presented in Chapter 9 – just one concordance from one corpus. There is nothing to suggest that the insights which have been gained from analysing this one concordance will not be applicable to others, even if other words show up other facets which enlarge upon the conclusions which we have been able to draw about the 'date' concordance. This is corroborated by the high degree of overlap demonstrated with a second concordance set, for the node word 'exchange'. Given the obvious, and necessary, limitations of the evaluation process, there is no guarantee that the system will work with all words and that the network of bonds might become overly complex for very frequent words with complex collocational profiles. In such instances, it might be profitable to see what improvement could be made by the integration of related analytical techniques such as lexical clustering (Phillips 1989) or text colonies (Rammell

1988).

As to the representativeness of the corpus, the BoE is certainly still among the largest in regular use for lexicographical purposes, but this may not always be the case, since there will always be a demand among some members of the corpus-using community for ever bigger corpora, additionally fuelled by the constantly growing capabilities of the hardware. This being the case, the application of filtering techniques such as the approach described herein will become more and more crucial to the successful exploitation of these ever-larger corpora. In terms of its content, BoE is likely to be as balanced as any other corpus of similar size, but it is interesting to dwell for a moment on the issue of corpus construction and the effects that this might have on the performance of the analytical software. It seems likely that a less balanced corpus, for example a genre-specific collection of texts, would exhibit less variation in the collocational patterning presented in the concordances, possibly resulting in concordance lines which contained a higher density of link words. This should not be a cause for concern, however, since by adjusting the parameters, it ought to be possible to accommodate data with a higher degree of linking, either through raising the link threshold or by incorporating new items into the stopword list.

The results of the evaluation would seem to indicate that the application of cohesive analysis is a useful means of providing such a filtering mechanism, since a high degree of correlation with the manual analyses was achieved for several sets of software parameters. Comparing the parameters with those used in automatic abridgement for a moment, it is interesting to note that the most successful stopword lists, in terms of how often they figured in highly-correlated parameter combinations, were those which contained fewest items. This is very different from the extensive stopword list employed in the abridgement process, which excludes all grammatical words, reinforcing the point made during the discussion of stopwords in Chapter 6, that the cohesion between concordance lines is operating at a level beyond that of the lexis alone and that something which might be

called 'contextual cohesion' is taking place between the individual lines of the concordance.

Apart from the hoped-for correlation between the human experts' and the computer's analyses, another striking result to come out of Chapter 9 was the marked differentiation between the parameter combinations required in order to identify representative lines and those needed for usable ones. The fact that a fixed span and a small or empty stopword list provided the best correlation with the respondents' selection of usable lines appears to indicate that the informants are making use of features which are more closely-defined positionally and heavier in grammatical items than those which occur in the lines which are chosen as representative. There are a number of lines of enquiry which might be pursued in order to investigate this dichotomy more closely, as well as other points relating to possible future enhancements to the *cohort* software. The final sections of this chapter will briefly touch on these.

10.2. Extending the Node Word

One variation on the system, which would be applicable to the further study of the distinction between usable and representative lines, would be to extend the concept of the concordance's node word. As noted in Chapter 3, the node word is automatically excluded from the link analysis, but it would also be possible to filter out particular grammatical words which were forming links. Suppose, for instance, that we were to examine only those lines in the 'date' concordance where 'to' occurred to the left of the node word, but that we also excluded the item 'to' from the link analysis. This approach would allow a more focused study to be made of any other features of this subset of lines which were contributing to link formation – in this case, the various phrases and other features which involve the item 'to' such as 'to date' (until now), 'to date' (infinitive) + *x*, 'to this/that date', 'up to date', 'keep up to date'. Of course, this method could also be applied to lexical items within the context of the node word, e.g. 'postponed', 'set' or

'fixed', but the results from the previous chapter suggest that this would probably identify representativeness rather than usability and would require one of the less strict positional definitions, in contrast to the grammatical items, which successfully form links even when the absolute link type is used, because they are being used in stereotypical constructions with a fixed positional relationship to the node word.

10.3. Flexible Positioning

The issue of positional specification leads into another possible line of enquiry. While there are phrases, composed of grammatical items, which are entirely fixed, there are others which are more free to vary, within fairly tight constraints, in their form. A simple example of this in the 'date' concordance would be the determiners, where it would be useful to know whether 'the' or 'a' is preferred. These items always occur (obviously) in the left-hand context and so would not be differentiated by the relative link type, while the absolute type would not allow the freedom of movement caused by the insertion of premodifiers between the determiner and the node word, as in 'set a date', 'set a new date' etc. In order to accommodate this variation, a more flexible approach to positional specification might be introduced, which divided the context into 'bands' several words wide rather than using single-word slots; links would then be formed between occurrences of any item in the same band. Alternatively, some kind of fuzzy match might be applied to the absolute positioning information, so that items could still be linked, even though they did not occur in exactly the same slot. This approach could be employed to identify any item which always occurs on the same side of the node word, but which is free to drift within one or two positions and so might be useful in identifying lexical links as well as grammatical ones.

10.4. Link Word Frequency

As an introduction to a further possible enhancement to the *cohort* software, let us refer back to the discussion of Figure 7.1, where it was noted that no parameter in the current software system has an effect on the transformation of the wordlist into the matrix. A possible extension to the software, which would bring about this transformation, would be the addition of a parameter which defined a minimum frequency for individual link words. This would actually be a progression from an existing rule in the program which states that a wordlist item must occur in at least two sentences (since it could not possibly form a link otherwise). By raising this threshold, some of the low-level links would be eliminated, removing some of the 'noise' associated with low frequency collocates and also enabling the software to execute more quickly, as fewer pairs of lines would need to be recorded in the matrix. Applying this to the 'date' concordance wordlist built with the raw link type, over half of the nearly 12,000 links would be eliminated if the frequency threshold were raised to five.

10.5. New Link Classes

It was noted in Section 9.8 that no attempt has been made to incorporate part-of-speech information into the cohesion analysis – this currently functions solely on the basis of the word string as it appears in the wordlist. If it were possible to obtain a syntactically-tagged version of the 'date' concordance, then the analyses could be repeated using a method which took into account the part-of-speech information. Given the concordance fragment below:

```
1) ad declared their willingness to set a date for starting stage two of economic an
2) ? Almost half thought she should set a date for stepping down; 35 per cent that s
3) pe given here is set for this time and date, and for the capital, Paramaribo.<t>
4) n Dublin on Saturday should set a firm date for an inter-governmental conference
5) d last night to set 2000 as the target date for stabilising emissions of carbon d
6) Board and set an implementation target date of January 1. <t> The working party h
7) pendence, and Chart 91 is set for this date for Helsinki, the capital, for 12.00
```

the entry for the word 'set' in the wordlist would normally just reflect the link type in effect, either 'set' or 'set:–' or for the absolute link type:

set:–2 1 2
set:–3 4 7
set:–4 6
set:–5 3 5

If word-class data were to be available, the wordlist information could be augmented, resulting in something like:

set_INFINITIVE 1 2 4 5
set_PAST-PARTICIPLE 3 7
set_SIMPLE-PAST 6

The positional information attached to the wordlist items could be retained (e.g. set_INFINITIVE:–2) and would serve its normal purpose of restricting the placement of the (word+)tag in relation to the node word. As an alternative, the word string could be suppressed completely and the entire process carried out entirely on the basis of the part-of-speech tags, thus identifying links on the basis of shared *grammatical* context.

This type of approach might enable us to shed more light on the usability/representativeness distinction. Given a set of lines in which strong bonding was identified when using parameters most strongly correlated with usability, it would be interesting to note the degree of overlap with a set of lines whose bonds were generated on the presence of particular repeated grammatical items using a similar set of parameters. The occurrence of a significant number of lines shared between the two sets would tend to support the proposition that the informants' choice of usable lines was indeed being influenced by the presence of grammatical items within the context of the node word.

The availability of part-of-speech or phrase structure information within the concordance would open up other attractive avenues of research. If the position of each word within the concordance could be defined in terms of its role within the sentence structure from which it was drawn, then the analysis of the concordance could be carried out along rather different lines. The presence of the metatextual tags would make it possible to

create links between conventional items in the concordance line and the metatextual data, so that, for example, a line could become linked because of the presence of the subject noun phrase within it or or because the node word was part of the right-hand context of an adjective. The links thus created would contribute to the overall bonding of each line in exactly the same fashion as we have seen previously. The result, however, would be a set of lines selected on the basis of the typicality of their grammatical construction, always supposing, naturally, that tools are available which will recognise the word class and phrase structure for each word in the corpus to a satisfactory degree of accuracy. The possibility exists, however, that the POS information could cause valid links across word classes to be lost. By looking for a certain number of link words within the context of the node, a degree of disambiguation is already carried out. This requires further investigation with POS-tagged data.

10.6. Putting the Software to Work

In conclusion, it is perhaps worth considering some of the contexts in which the analytical approach described herein might be brought to bear on the problems associated with the use of today's large-scale corpus resources. It is undeniably true that the users of these corpora need more sophisticated tools: as we saw in Chapter 5, Sinclair and Clear felt the need to develop their *typical* program even though their corpus amounted to only twenty million words at the time, while in Chapter 2 Biber pointed out the overwhelming amount of evidence produced by a mere ten million words, yet these corpus sizes are only a fraction of the amount of material which is now available to corpus researchers. The approach I have described is not intended to stand alone as *the* means of dealing with the corpus information overload, but rather it should be seen as a prelude to normal corpus inspection, so that other tools such as sorting and collocate analysis, described in Chapter 2, would be applied once the number of lines had been reduced to more manageable dimensions. Ideally, the analytical engine of *cohort* would be incorporated into a suite of

corpus software, as this would introduce even greater flexibility in the choice of span size, the use of word-class tags, sentence-length concordances and so on.

The computer hardware capable of accessing large-scale corpus resources is already widely available, but the technological advances which have enabled the creation and distribution of such extensive collections of electronic text must be mirrored by the development of software that is able to exploit those resources to their full potential. This will ultimately be of benefit not just to a small group of professional users of bespoke corpora but also to the growing number of researchers and teachers who are making use of any of the available corpus resources.

Bibliography

Collier

References

Benbrahim, Mohamed and Khurshid Ahmad (1995) 'Text summarisation: the role of lexical cohesion analysis' in *The New Review of Document and Text Management*, Volume 1, 1995. Taylor Graham, London, pp 321-335.

Biber, Doug (1993) 'Representativeness in corpus design' in *Literary and Linguistic Computing* 8(4) pp 243-257.

Biber, Doug, S. Conrad and R. Reppen (1994) 'Corpus-base approaches to issues in applied linguistics' in *Applied Linguistics* 15(2) pp 169-189.

Brill, Eric (1992) 'A simple rule-based part-of-speech tagger' in *Proceedings of the Third Conference on Applied NLP*, Trento, Italy.

Brown, Peter F., Peter V. de Souza, Robert L. Mercer, Vincent J Della Pietra and Jenifer C. Lai (1992), 'Class-Based *n*-gram Models of Natural Language' in *Computational Linguistics*, volume 18, number 4, ACL, MIT Press, pp 467-479.

Burnard, Lou (ed) (1995) *Users Reference Guide for the British National Corpus*. Oxford University Computing Services, Oxford.

Butler, Christopher S. (1985) *Statistics in linguistics*. Blackwell, Oxford.

Church, Ken, William Gale, Patrick Hanks and Donald Hindle (1990) 'Using Statistics in Lexical Analysis' in *Lexical Acquisition: Using On-line Resources to Build a Lexicon*, Zernik, U. (1990) (Ed.) Lawrence Erlbaum Associates, New Jersey.

Church, Ken & Patrick Hanks (1989) 'Word association norms, mutual information and lexicography' in *Proceedings of the 27th Annual Meeting of the ACL, Vancouver, Canada, pp 76-83.*

Clear, Jeremy H (1987) 'Computing' in Sinclair (1987) pp 41-61.

Clear, Jeremy H (1988) 'Trawling the Language: Monitor Corpora' in M. Snell-Hornby (ed) (1987) *ZüriLEX '86 Proceedings*. Francke, Tübingen.

Clear, Jeremy H (1995) '"Grammar and nonsense": or syntax and word senses' in *Words: Proceedings of an International symposium*, Lund 25-26 August 1995, J. Svartvik (ed). Kungl. Vitterhets Historie och Antikvitets Akademien, Konferenser 36, Stockholm.

Collier, Alex (1993) 'Issues of Large-scale Collocational Analysis', in *English Language Corpora: Design, Analysis and Exploitation*, Aarts, J, P de Haan & N Oostdijk (eds) (1993). Rodopi, Amsterdam.

Collier, Alex (1994) 'A system for automating concordance line selection', in *Proceedings of the International Conference on New Methods in Language Processing*, CCL, UMIST, pp 95-100.

Collier, Alex with Antoinette Renouf (1995) 'A system of automatic textual abridgement', in *Proceedings of the 15th International Conference on Language Engineering*, Montpellier, 27-30 June 1995.

Collier, Alex & Mike Pacey (1996) 'A Large-scale Corpus System for Identifying Thesaural Relations', in Ljung, Magnus (ed) *15th International ICAME Conference Proceedings*. Rodopi, Amsterdam.

Cruden, Alexander (1828) *complete concordance to the Holy Scriptures of the Old and New Testament'*. C & J Rivington, London.

Daille, Béatrice (1995) 'Combined Approach for Terminology Extraction: Lexical Statistics and Linguistic Filtering', UCREL Technical Report 1995:5, University of Lancaster.

Even-Shoshan, Abraham (1977) *A new concordance of the Bible*. Kiryat Sapher Publishing House, Jerusalem.

Firth, JR (1957) 'A Synopsis of Linguistic Theory 1930-1955' in F. Palmer (ed): *Selected Papers of JR Firth*. Longman, London.

Garside, Roger (1987) 'The CLAWS word-tagging system', in Garside, Leech and Sampson 1987.

Garside, Roger, G Leech & G Sampson (eds) (1987) *The Computational Analysis of English: A Corpus-based Approach*. Longman, London.

Halliday, Michael A.K. & Ruqaiya Hasan (1976) *Cohesion in English*. Longman, London.

Hanks, Patrick (1987) 'Definitions and Explanations' in Sinclair (1987) pp 116-136.

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

Hoey, Michael (1997) 'From Concordance to Text Structure: New Uses for Computer Corpora' in *Proceedings of PALC'97*, University of Lodz, Poland, 12-14 April 1997, pp 2-23.

Hofland, Knut & Stig Johansson (1982) *Word Frequencies in British and American English*. Longman, Harlow.

Johnson, Samuel (1755) *A Dictionary of the English Language*. Facsimile of 1755 edition published in 2 volumes. Times Books, London, 1979.

Källgren, Gunnel (1988) 'Automatic Abstracting of Content in Text' in *Nordic Journal of Linguistics*, 11, pp 89-110.

Kaufman, Leonard & Peter Rousseeuw (1990) *Finding Groups in Data*, Wiley Series in Probability and Mathematical Statistics. John Wiley and Sons, London.

Kucera, Henry & W. Nelson Francis (1967) *The Computational analysis of present-day American English*. Brown University Press, Providence, R.I.

Leech G.N. (1966) *English in Advertising: a linguistic study of advertising in Great Britain*. Longman, London.

McCarren, V.P. (1977) *A critical concordance to Catullus*. Brill.

McEnery, Tony and Andrew Wilson (1996) *Corpus Linguistics*. Edinburgh University Press, Edinburgh.

Collier

Marco, Maria Jose Luzon (1999) 'Procedural Vocabulary: Lexical Signalling of Conceptual Relations in Discourse' in *Applied Linguistics*, 20(1), 1999, pp 1-21.

Morris J and G Hirst (1991) 'Lexical Cohesion computed by Thesaural Relations as an Indicator of the Structure of the Text' in *Journal of Computational Linguistics*, 17(1), 1991, pp 21-48.

Murray, Lindley (1851) *English grammar adapted to the different classes of learners with an appendix containing rules and observations for assisting the more advanced students to write with perspicuity and accuracy* 57th edition. Longman, London.

Phillips, Martin (1989) *Lexical Structure of Text*. University of Birmingham: English Language Research.

Rammell, Christina (1988) 'A non-literary stylistic analysis of the statute and its internal organisation', MA project, University of Birmingham.

Reimer, U and U Hahn (1990) 'An Overview of the Text Understanding System TOPIC' in Ulrich, Scmitz, Rüdiger Schütz and Andreas Kunz (eds) *Linguistic Approaches to Artificial Intelligence*. Verlag Peter Lang, Frankfurt, 1990.

Renouf, Antoinette (1986) 'Lexical Resolution' in Meijs, W (ed) *Corpus Linguistics and Beyond: The proceedings of the 7th International Conference of English Language Research on Computerised Corpora*. Rodopi, Amsterdam.

Renouf, Antoinette (1987) 'Corpus Development' in Sinclair (1987) pp 1-40.

Renouf, Antoinette (1992) 'The AVIATOR Project' in *Proceedings of the 11th International ICAME Conference*. Rodopi, Amsterdam.

Renouf, Antoinette (1993) 'A word in time: First findings from the investigation of dynamic text' in *English Language Corpora: Design, Analysis and Exploitation*, Aarts, J, P de Haan & N Oostdijk (eds) (1993). Rodopi, Amsterdam.

Renouf, Antoinette (1996) 'The ACRONYM Project: Discovering the textual thesaurus' in *Synchronic corpus linguistics: Papers from the 16th International Conference on*

English Language Research on Computerized Corpora in Toronto, Percy, C, C Meyer & I Lancashire (eds) (1996). Rodopi, Amsterdam.

Salton, G, J Allan, C Buckley and A Singhal (1994) 'Automatic Analysis, Theme Generation and Summarisation of Machine Readable Texts' in *Science*, 264(3), 1994, pp 1421-1426.

Schütze, Hinrich and Jan Pedersen (1993) 'A vector model for syntagmatic and paradigmatic relatedness' in *Proceedings of the 9th Annual Conference of the UW Centre for the New OED and Text Research*, pp. 104-113, Oxford, England.

Sinclair, John M. (ed) (1987) *Looking Up: An account of the COBUILD Project in Lexical Computing*. Collins, London.

Sinclair, John M. (ed) (1987a) *Collins-Cobuild English Dictionary*. Collins, London.

Sinclair, John M. (ed) (1987b) 'Collocation: a progress report' in Steele R and T Threadgold (eds) *Language Topics: an internation collection of papers by colleagues, students and admirers of Professor Michael Halliday to honour him on his retirement* Volume 2 pp 319-331. John Benjamins, Amsterdam.

Sinclair, John M. (1990) *Collins Cobuild English Grammar*. HarperCollins, London.

Sinclair, John M. (1991) *Corpus, Concordance, Collocation*. Oxford University Press, Oxford.

Stenström, Anna-Brita & V. Haslerud (1995) 'Transcribing COLT: The Bergen Corpus of London Teenage Talk' in G. Leech, G. Myers & J. Thomas (eds) *Spoken English on Computer*, pp 235-242. Longman, London.

Stolcke, Andreas and Jonathan Segal (1994) 'Precise *n*-gram Probabilities from Stochastic Context-free Grammars' in *Proceedings of the Annual Conference of the ACL*, 1994.

Stubbs, Michael (1995) 'Collocations and Semantic Profiles: On the Cause of the Trouble with Quantative Studies' in *Functions of Language 2, 1: 23-55*, John Benjamins BV,

Amsterdam.

Svartvik, Jan (ed) (1990) *The London-Lund Corpus of Spoken English: Description and research*. Lund Studies in English 82. Lund University Press, Lund.

Svartvik, Jan (1991) 'Corpora are becoming mainstream' in Thomas & Short pp 3-13.

Thomas, Jenny & Mick Short (eds) (1996). *Using Corpora for Language Research: Studies in Honour of Geoffrey Leech*. Longman, London.

Voutilainen, Atro & J Heikkilä (1994) 'An English Constraint Grammar (EngCG): a surface-syntactic parser of English' in Fries, Tottie & Schneider (eds) (1994) *Creating and Using English Language Corpora*. Rodopi, Amsterdam.

Wangheng, Peng (1998) 'Keywords in Theme and their implications for TEFL reading', presented at Teaching and Language Corpora 1998 (TALC 98), July 24-28, Keble College, Oxford.

Glossary

Glossary

100M A corpus built by the author at the University of Birmingham in the late 1980s. It contained BCET and a collection of British national daily newspapers, amounting to just under 100 million tokens. It was mainly used to test corpus retrieval software and verify projections of corpus growth.

ACR The corpus database which was utilised in the SERC-DTI funded ACRONYM Project at the University of Liverpool (1994-97). It was entirely composed of newspaper data (Independent and Financial Times) and at the end of the project totalled approximately 435 million tokens.

Bank of English or **BoE** A balanced collection of corpus data in excess of two hundred million words in size. Containing speech and written material, it was built by HarperCollins Publishers in the mid 1990s in order to replace the BCET (*qv*) as the basis for the EFL reference works produced by Cobuild.

BCET The Birmingham Collection of English Texts. A twenty-million word corpus of 25% speech/ 75% writing established at the University of Birmingham in the early 1980s and used in the first Cobuild dictionaries and other reference works until superseded by the Bank of English. See Renouf 1987.

British National Corpus or **BNC**. A balanced corpus funded as a SERC-DTI project. One hundred million words in extent, tagged for part of speech and made available on CDROM in SGML format. Produced through collaboration of the Universities of Oxford and Lancaster and Longman publishers.

Brown The Brown Corpus. A one-million word corpus of samples of text taken from a selection of genres of American printed matter, built at Brown University in the 1960s. See Kucera and Francis 1967.

Collocational density The relationship between the number of times a word occurs and the number of significant collocates it has.

LOB The London Oslo Bergen Corpus, built on the same model as the American *Brown* corpus (*qv*). See Hofland & Johansson 1982.

London-Lund A 500,000-word corpus of spoken material, built in collaboration between UCL and Lund University. See Svartvik 1990.

parser A piece of software which analyses the phrase structure of a corpus text (which will often have been previously passed through a tagger) and produces a version of the text which is marked-up accordingly.

tagger A piece of software which automatically assigns part-of-speech labels (tags) to each word in a corpus text.

token A single word in a corpus. Also called 'running word'. Generally used in measurements of corpus size. Cf *type*.

tokeniser The part of a software program which breaks a text into individual words or tokens, separating them out from punctuation, white space etc and attaching any extra information such as the current sentence number.

type or **word type** A discrete word form

wordlist The list of all the word types which have occurred in a set of concordances. After each word type are listed the line numbers of the concordance lines in which they were found.

Appendix 1
Stopword Lists

arts-prons

a	an	and	he	her	his
i	it	she	the	they	this
we	which	you			

bt

a	about	above	across	after	against
all	along	alongside	also	although	always
am	amid	amidst	among	amongst	an
and	any	anybody	anyone	anything	anywhere
apropos	are	aren't	as	at	atop
be	because	been	before	behind	being
below	beneath	beside	besides	between	beyond
both	but	by	can	can't	cannot
cos	could	couldn't	coz	dare	daren't
despite	did	didn't	do	does	doesn't
doing	don't	done	dr	during	each
either	else	every	everybody	everyone	everything
everywhere	except	few	for	from	go
going	had	hadn't	has	hasn't	have
haven't	having	he	he'd	he'll	he's
her	here	hers	herself	him	himself
his	how	however	if	in	inside
into	is	isn't	it	it'd	it'll
its	itself	less	like	make	many
may	mayn't	me	might	mine	minus
more	most	mr	mrs	much	must
mustn't	my	myself	needn't	neither	never
nevertheless	no	no-one	nobody	none	nonetheless
noone	nor	not	nothing	notwithstanding	of
off	often	on	one	only	or
other	ought	oughtn't	our	ours	ourselves
out	outside	over	part	per	plus
rather	shall	shan't	she	she'd	she'll
she's	should	shouldn't	since	so	some
somebody	someone	someplace	something	sometime	sometimes
somewhere	than	that	that'd	that'll	that's
the	thee	their	theirs	them	themselves
then	there	there'd	there'll	there's	there've
therefore	therewith	these	they	they'd	they'll
they're	they've	thine	this	those	thou
though	through	throughout	thus	thy	till
to	too	toward	towards	under	underneath
until	up	upon	us	very	via
was	wasn't	well	were	what	what'd
what'll	what's	what've	whatever	when	whenever
where	wherever	whether	which	whichever	while
whilst	who	whom	whose	why	will
with	within	without	won't	would	wouldn't
ye	yeah	yes	you	you'd	you'll

you're	you've	your	yours	yourself	yourselves

btb

a	all	also	although	always	am
an	and	any	anybody	anyone	anything
anywhere	apropos	are	aren't	as	be
because	been	being	besides	both	but
by	can	can't	cannot	cos	could
couldn't	coz	dare	daren't	did	didn't
do	does	doesn't	doing	don't	done
dr	each	either	else	every	everybody
everyone	everything	everywhere	few	go	going
had	hadn't	has	hasn't	have	haven't
having	he	he'd	he'll	he's	her
here	hers	herself	him	himself	his
how	however	if	is	isn't	it
it'd	it'll	its	itself	less	make
many	may	mayn't	me	might	mine
minus	more	most	mr	mrs	much
must	mustn't	my	myself	needn't	neither
never	nevertheless	no	no-one	nobody	none
nonetheless	noone	nor	not	nothing	of
often	one	only	or	other	ought
oughtn't	our	ours	ourselves	per	plus
rather	shall	shan't	she	she'd	she'll
she's	should	shouldn't	so	some	somebody
someone	someplace	something	sometime	sometimes	somewhere
than	that	that'd	that'll	that's	the
thee	their	theirs	them	themselves	then
there	there'd	there'll	there's	there've	therefore
therewith	these	they	they'd	they'll	they're
they've	thine	this	those	thou	though
thus	thy	to	too	us	very
was	wasn't	well	were	what	what'd
what'll	what's	what've	whatever	when	whenever
where	wherever	whether	which	whichever	while
whilst	who	whom	whose	why	will
with	within	without	won't	would	wouldn't
ye	yeah	yes	you	you'd	you'll
you're	you've	your	yours	yourself	yourselves

top50

a	all	an	and	are	as
at	be	been	but	by	for
from	had	have	he	her	his
i	if	in	is	it	my
no	not	of	on	one	or
p	said	she	so	that	the
their	there	they	this	to	was
we	were	what	when	which	with
would	you				

top100

a	about	after	all	an	and
any	are	as	at	back	be
because	been	but	by	can	could
did	do	don't	down	even	first
for	from	get	had	has	have
he	her	him	his	how	i
if	in	into	is	it	its
just	know	like	man	me	more
most	much	my	no	not	now
of	on	one	only	or	other
our	out	over	p	people	said
see	she	so	some	than	that
the	their	them	then	there	these
they	think	this	time	to	two
up	very	was	way	we	well
were	what	when	which	who	will
with	would	you	your		

top150

a	about	after	again	all	also
an	and	another	any	are	as
at	away	back	be	because	been
before	being	between	but	by	came
can	come	could	day	did	do
don't	down	even	first	for	from
get	go	going	good	got	had
has	have	he	her	here	him
his	how	i	if	in	into
is	it	it's	its	just	know
life	like	little	long	made	make
man	many	may	me	might	more
most	much	must	my	never	new
no	not	now	of	off	old
on	one	only	or	other	our
out	over	own	p	people	put
right	said	same	say	see	she
should	so	some	something	still	such
take	than	that	the	their	them
then	there	these	they	think	this
those	thought	through	time	to	too
two	up	us	very	was	way
we	well	went	were	what	when
where	which	who	will	with	work
world	would	years	yes	you	your

zero

This contain no stopwords.

Appendix 2
Concordance of 'exchange' from BCET

001 Condorcet in Paris under an "au pair" exchange when two French children came
002 Ruder, chairman of the Securities and Exchange Commission, said Britain, the
003 a no-fault situation. I advise you to exchange names and be on your own way.
004 and see me again. It'll do us good to exchange ideas." She could have been gl
005 came from Berlin and abroad, eager to exchange the new ideas that were racing
006 could utter silence. In practice, the exchange of letters often takes a full
007 endeavoured to heal the wounds. In an exchange of letters with Mansholt he de
008 exchanges; genes within bacteria can exchange. But, in the past at least, it
009 foreign services usually press for an exchange, and often in poor countries t
010 he found him to Miss Gray at the Corn Exchange, where he would be suitably re
011 here was unbearable. And he wanted to exchange the unbearable for the very ba
012 home Britain already has one System X exchange working, and should have eight
013 horses, beads and cloth came south in exchange. These societies were so far f
014 it ahead of her, how it would be. The exchange of witty letters, fewer as tim
015 know where they gave the best rate of exchange. The whole place was reflected
016 market-place. There, outside the Corn Exchange which dominates one side of th
017 new house (the 10 per cent deposit on exchange of contracts) before you've re
018 of value, the proposition that things exchange in accordance with the amount
019 require garages where you can simply exchange a battery pack for a fully cha
020 smiled and again indicated the corner exchange. "The bridge phone hangs just
021 solution would be for it to lower its exchange rate vis-a-vis other countries
022 that there had been some way he could exchange words with this man he had nev
023 the best way in an emergency. The SIS exchange called Boyd Stuart's home and
024 them (as used to be said on the Stock Exchange), cast no doubt envious glance
025 to freer, cheaper and more widespread exchange of information between the ric
026 to recapture that lead with a digital exchange called System X. It is not goi
027 two marine insurance firms, the Royal Exchange Insurance and the London Assur
028 when its shares are introduced to the Exchange, probably in January. He also
029 which we have witnessed on the stock exchange this week, does the team agree
030 why the stranger hadn't completed the exchange of names like any other decent
031 , because he hadn't a card to offer in exchange. Mothersole, he could feel, wa
032 , price controls and food subsidies in exchange for voluntary wage restraint,
033 - Because (Nora?) told me about school exchange and I went along to see for m
034 And priests were extracting "gifts" in exchange for burying non-churchgoers in
035 CF Yes. RB Due to import licencing or, exchange er control regulations. CF Uhm
036 Copies were burned on the London Stock Exchange and destroyed at exchanges in
037 Flanders. The printer was entranced to exchange a few of the place-names which
038 Lou Darrow Carrington runs the foreign exchange desk for the bank's corporate
039 MALA TO GIVE UP ITS CLAIM TO BELIZE IN EXCHANGE FOR CERTAIN ECONOMIC CONCESSIO
040 N A SINGLE IMAGINATIVE GESTURE. AT THE EXCHANGE WE GET THE OLD SPECTACLE OF A
041 ONAL. IT"S A RATHER INTENSE PIECE, AN EXCHANGE BETWEEN TWO PEOPLE, AND IT"S A
042 Thomas traveled inside; but he did not exchange two remarks with Fanny during
043 W YORK - Prices on the New York Stock Exchange staged a blue- chip rally F
044 a and are the same kind of judgemental exchange these ladies, as children, ove
045 a for them, and I keep him supplied in exchange for plenty fires and troubles
046 a our Unit Trusts then we have a Share Exchange Scheme whereby you can obtain
047 aily basis and there wll always be an exchange risk. <P 5> ((How to open an
048 al - and end up with failed degrees in exchange for a phenomenal understanding
049 al Festival, which is held in the Corn Exchange each May. Miss Gray and I had
050 all of us if we did not calm down. Our exchange was heated. Within a matter of
051 ambling clubs or drinking at the Royal Exchange Pub and talking about politics
052 amental right to adequate treatment in exchange for being deprived of his libe

053 and maintain the correct rate of fluid exchange within the various fluid compa
054 and manufactures to the third world in exchange for raw materials and food, is
055 and other statues from the first Royal Exchange). The Library is a first-class
056 ange of biographical information. This exchange strengthens the bnd because i
057 any. Often their beginning has been an exchange of correspondence between <P 1
058 ar's nominee for the Rummidge-Euphoria exchange scheme. Why should Morris Zapp
059 at was not old at all. "The Merchants" Exchange is a ruin . . . is that pleasa
060 ate a small plot on the worst land, in exchange for agricultural and even dome
061 ay from each other. As they talk, they exchange only the briefest of glances.
062 blushing, as the following therapeutic exchange demonstrates: Therapist: "Why
063 ce economies (where very little market exchange takes place), this form is on
064 conomy changed from one based upon the exchange of goods and services to one b
065 currency accounts was established when Exchange Control Regulations were lifte
066 d an expert in Round Tableip and the Exchange of Unpleasantries. Last certai
067 d raw materials naturally need foreign exchange to buy these, but because of t
068 d, "is not the way to begin a cultural exchange." The incident caused the trai
069 der our range of Unit Trusts. ((Share Exchange.)) If you already own some st
070 ders shoot up? Sometimes a good verbal exchange and a bit of arm waving will d
071 ds over her identity to her husband in exchange for a small portion of his, sh
072 e between the target telephone and the exchange) are more advisable- There are
073 e disposed of, and offered for sale in exchange for cash - and when cash is no
074 e to understand what they say and even exchange complex messages with them. So
075 ecting <P 226> investment and imposing exchange controls. I ondered how he pr
076 ee market, for which you need an equal exchange between equal parties. Even wh
077 ehall but best remembered by me for an exchange of Jack Buchanan's signed ciga
078 elt like a Dwarf - Like me - Was he an exchange, Betty, is Alan going? - No, A
079 eneral impression. He was moved by the exchange of vows, the old clear words,
080 er was an indeterminate little man. In exchange for our money, they were suppo
081 f papers. "Well, there is the Rummidge exchange, but you wouldn't be intereste
082 ferent workplaces and jobs can meet to exchange their experiences. In other wo
083 for how they would live. She would not exchange her solitude for anything. Nev
084 for the walls of their private bars in exchange for a few pints of beer. Or, i
085 get a quicker reactionfrom the Berlin Exchange. The exchange would be told to
086 ght to harvest any farmer's fields. In exchange, they get one ninth of the cro
087 gned to work with the local service to exchange information, to train the loca
088 h it's not legally binding). The Stock Exchange itself requires any public com
089 hat genes from bacteria don't normally exchange with genes from humans or gen
090 hat is, to establish the exact rate of exchange at which mechanical energy is
091 hat the offence might be overlooked in exchange for a consideration: they woul
092 he flat and go to a hotel. The rate of exchange in Denmark is heavily against
093 hem absolutely and Alan's supposed to exchange back there? - Yes, well as I s
094 hey would be given decent treatment in exchange for "honest labor." ZOB iasued
095 hip, instant engagement, and immediate exchange of marriage vows. Thus it has
096 his beneficence I drew fro the Labour Exchange for six months had to be eked
097 id enjoyment. The justification is the exchange of ideas, and the value of thi
098 imise our clients' exposure in foreign exchange. We tell them what's happening
099 inder of the session was devoted to an exchange on the compatibility of religi
100 ist, the business tycoon (on the Stock Exchange), two mathematicians, three fa
101 kbrokers did by the pillars of the old Exchange. Through the west gate of this
102 ld be impossible for the "fighters" to exchange roles so freely. They would be
103 llowing questions were overheard in an exchange between two five-year-olds: Wo
104 ltivation can be surprisingly high, in exchange for no investment in fertilize
105 ly escalation into a strategic nuclear exchange between the Soviet Unon and t
106 ly the Young Communist League has made exchange visits with Komsomol, the Sovi

man gives food, care and protection in exchange for the different services the
markets, with its own model of digital exchange as its standard-bearer, comple
mber of units in any of our Trusts in exchange for your securities - this exc
me had become an international rate of exchange. As he wandered through the st
me up," Dixon said. He settled a brief exchange with the taxidiver by refusing
means of production, distribution, and exchange is profitability; that any dep
means of production, distribution, and exchange", meant something quite differ
means of production, distribution, and exchange". The prose style of the notor
mena's trip has sparked a sharxpublic exchange between him and Velasco. The F
mmals. Even a relatively small nuclear exchange would cause severe after-effec
mpete if we are to earn enough foreign exchange to buy the primary goods we st
n and averts severe strains on foreign exchange. A nation with stockpiled rese
n they were seen together on the Stock Exchange floor, are joining forces agai
n which every shared bite is a sensual exchange. So I said again, "Come on, no
nch is a forum which allows members to exchange views and opinions. It is also
nd ack- nowledged him as their Lord in exchange for whatever they most desired
ne was out of order; on the second the exchange was closed for a religious hol
ng deposits). <P 8> As with all Stock Exchange investment price can go down
ng to hand back all their conquests in exchange for the tiny border enclaves
ng unless you're fiddling on the Stock Exchange, which I don't recommend. It's
nician at the city's central telephone exchange, and that he had often tried t
nso Lopez Michelson, to visit Quito to exchange experiences and to promote PLP
ocal shop. We can give you sterling in exchange for most foreign notes but coi
oco had reported to the Securities and Exchange Commission. Recently, Texaco l
olas Goodison, ? Chairman of the Stock Exchange, was asked if he found the lar
ome s"pose" and so on. There is a sure exchange :CHANGED of thought and some p
onsequently, er you know, there's free exchange in that way - mmm - though
orated unconvincingly intoa telephone exchange). The headquarters of the Anti
ould start to fall on the foreign exchange markets, particularly again
ound 80. Bear in mind that the rate of exchange while 1 was there was 11.20 fr
ous abbreviation on the New York Stock Exchange. A stock listed as Spud appare
ouse production a main item of foreign exchange or moneyearning, only the simp
own by the passenger's window. A brief exchange. Hogan waved, clipping on a se
papers presented to the Securities and Exchange Com- mission, the multinationa
pawn your land for five years or so in exchange for the cash. The moneylender
pproached the defense table, hoping to exchange a few words with them. The gu
pt to minimize the risk of a strategic exchange with the Soviet Unio~ with th
r thought of applying for the Euphoria exchange: "Not really, Gordon. It would
re businessmen of various sorts met to exchange money, property or goods. Afte
re of the enormous destruction such an exchange would cause, and this awarenes
responded Jimmy. Mrs Waites heard this exchange, and was torn between pity and
riday. You cannot cash a bank draft or exchange foreign currency when the bank
rille at Calcutta's central employment exchange, which must have one of the la
rld have a chance to meet together and exchange ideas. The Vegetarian Federal
rly since the introduction of floating exchange rates its become more importan
rovides a national focal point for the exchange of information, ideas and expe
s of bacteria and mammalian gnes may exchange in nature. So it may be that
sion. Another sent to me by the Labour Exchange presumably out of sheer kindne
son operations and their purpose is to exchange information, mount joint opera
t bearing deposits). As with all Stock Exchange investment prices can go down
t is that, apart from public telephone exchange equipment, PABXs are where the
te set for central bank money, but the exchange rate. The critical test is tha
tein, but it would only precipitate an exchange of feelings on a subject which
the cool under-water gloom of the Corn Exchange. Thankfully, I realized that t

161 the night officer and the sister would exchange a few words with us. In my fir
162 then next time I'll be the mamma. This exchange is not Child-Chld because of
163 tinued to a imported, and there was an exchange of "light" North Sea oil for "
164 trating digressions from the rewarding exchange of ideas he could enjoy with c
165 tries will fail to earn enough foreign exchange to maintain our primary base t
166 unting, president of the Toronto Stock Exchange. Mr Walker acknowleged that t
167 ush to sell. This hit the floor of the Exchange with torrential force. The mac
168 ut escalation into a strategic nuclear exchange. On the other hand, an observe
169 ve better than the USA after a nuclear exchange. Besides, there were enough wa
170 which recently received its full stock exchange listing. Debbie, aged 22, deci
171 who had received scholarships from the exchange program were given a trip to B
172 wine later, and frolicketwith an oral exchange of that, laughing over the hyd
173 y another name. In the absence of real exchange controls, however, the tax aut
174 y. Its for people who see that foreign exchange markets affect the things they
175 year. Beaverbrook glorified the formal exchange of letters between himself and
176 zzled people since 1881 : that when we exchange signals, then we discover that

Appendix 3
Parameter Combinations

Link Threshold	Link Type	Span Size	Stopword List	Number of Link Words	Total Links	Number of Bonded Lines	Total Bonds	St Dev
1	Abs	4	arts-prons	135	437	173	1960	8.34266
2	Abs	4	arts-prons	135	437	74	522	6.43428
3	Abs	4	arts-prons	135	437	21	26	0.447214
4	Abs	4	arts-prons	135	437	4	4	0
5	Abs	4	arts-prons	135	437	4	4	0
6	Abs	4	arts-prons	135	437	4	4	0
1	Abs	4	bt	51	138	98	410	4.31277
2	Abs	4	bt	51	138	16	20	0.447214
3	Abs	4	bt	51	138	4	4	0
4	Abs	4	bt	51	138	2	2	0
5	Abs	4	bt	51	138	0	0	0
6	Abs	4	bt	51	138	0	0	0
1	Abs	4	btb	73	238	134	1092	8.53815
2	Abs	4	btb	73	238	44	462	7.18331
3	Abs	4	btb	73	238	10	10	0.316228
4	Abs	4	btb	73	238	2	2	0
5	Abs	4	btb	73	238	0	0	0
6	Abs	4	btb	73	238	0	0	0
1	Abs	4	top100	51	139	99	410	4.15933
2	Abs	4	top100	51	139	18	22	0.547723
3	Abs	4	top100	51	139	6	6	0.316228
4	Abs	4	top100	51	139	2	2	0
5	Abs	4	top100	51	139	0	0	0
6	Abs	4	top100	51	139	0	0	0
1	Abs	4	top150	43	119	88	376	4.3359
2	Abs	4	top150	43	119	18	22	0.547723
3	Abs	4	top150	43	119	2	2	0
4	Abs	4	top150	43	119	2	2	0
5	Abs	4	top150	43	119	0	0	0
6	Abs	4	top150	43	119	0	0	0
1	Abs	4	top50	67	172	112	440	4.01248
2	Abs	4	top50	67	172	20	24	0.447214
3	Abs	4	top50	67	172	8	8	0.316228
4	Abs	4	top50	67	172	4	4	0
5	Abs	4	top50	67	172	0	0	0
6	Abs	4	top50	67	172	0	0	0
1	Abs	4	zero	177	690	175	4018	10.7564
2	Abs	4	zero	177	690	127	792	6.29285
3	Abs	4	zero	177	690	59	100	1
4	Abs	4	zero	177	690	16	20	0.316228
5	Abs	4	zero	177	690	6	6	0
6	Abs	4	zero	177	690	4	4	0
1	Abs	open	arts-prons	197	598	174	2292	8.68332
2	Abs	open	arts-prons	197	598	83	540	6.43428
3	Abs	open	arts-prons	197	598	27	34	0.447214
4	Abs	open	arts-prons	197	598	11	14	0.316228
5	Abs	open	arts-prons	197	598	4	4	0
6	Abs	open	arts-prons	197	598	4	4	0

Link Threshold	Link Type	Span Size	Stopword List	Number of Link Words	Total Links	Number of Bonded Lines	Total Bonds	St Dev
1	Abs	open	bt	76	193	112	470	4.34741
2	Abs	open	bt	76	193	16	20	0.447214
3	Abs	open	bt	76	193	11	14	0.447214
4	Abs	open	bt	76	193	4	4	0
5	Abs	open	bt	76	193	0	0	0
6	Abs	open	bt	76	193	0	0	0
1	Abs	open	btb	107	317	145	1194	8.61394
2	Abs	open	btb	107	317	45	464	6.97854
3	Abs	open	btb	107	317	15	18	0.447214
4	Abs	open	btb	107	317	8	8	0.316228
5	Abs	open	btb	107	317	0	0	0
6	Abs	open	btb	107	317	0	0	0
1	Abs	open	top100	73	186	111	456	4.15933
2	Abs	open	top100	73	186	18	22	0.447214
3	Abs	open	top100	73	186	11	14	0.447214
4	Abs	open	top100	73	186	4	4	0
5	Abs	open	top100	73	186	0	0	0
6	Abs	open	top100	73	186	0	0	0
1	Abs	open	top150	64	164	102	420	4.28952
2	Abs	open	top150	64	164	18	22	0.447214
3	Abs	open	top150	64	164	9	12	0.447214
4	Abs	open	top150	64	164	2	2	0
5	Abs	open	top150	64	164	0	0	0
6	Abs	open	top150	64	164	0	0	0
1	Abs	open	top50	93	227	122	492	4.03733
2	Abs	open	top50	93	227	20	24	0.447214
3	Abs	open	top50	93	227	11	14	0.447214
4	Abs	open	top50	93	227	6	6	0
5	Abs	open	top50	93	227	2	2	0
6	Abs	open	top50	93	227	2	2	0
1	Abs	open	zero	265	944	176	4590	11.2339
2	Abs	open	zero	265	944	140	888	6.38749
3	Abs	open	zero	265	944	63	114	1.09545
4	Abs	open	zero	265	944	19	24	0.447214
5	Abs	open	zero	265	944	11	14	0.316228
6	Abs	open	zero	265	944	6	6	0
1	Raw	4	arts-prons	140	638	176	7100	21.9431
2	Raw	4	arts-prons	140	638	132	894	7.34847
3	Raw	4	arts-prons	140	638	37	52	0.632456
4	Raw	4	arts-prons	140	638	4	4	0
5	Raw	4	arts-prons	140	638	4	4	0
6	Raw	4	arts-prons	140	638	2	2	0
1	Raw	4	bt	83	233	132	682	4.42719
2	Raw	4	bt	83	233	25	30	0.547723
3	Raw	4	bt	83	233	4	4	0
4	Raw	4	bt	83	233	2	2	0
5	Raw	4	bt	83	233	0	0	0
6	Raw	4	bt	83	233	0	0	0
1	Raw	4	btb	92	345	159	2910	17.0646
2	Raw	4	btb	92	345	61	534	7.18331
3	Raw	4	btb	92	345	10	10	0.316228

Link Threshold	Link Type	Span Size	Stopword List	Number of Link Words	Total Links	Number of Bonded Lines	Total Bonds	St Dev
4	Raw	4	btb	92	345	2	2	0
5	Raw	4	btb	92	345	0	0	0
6	Raw	4	btb	92	345	0	0	0
1	Raw	4	top100	84	236	134	710	4.32435
2	Raw	4	top100	84	236	27	32	0.547723
3	Raw	4	top100	84	236	6	6	0
4	Raw	4	top100	84	236	2	2	0
5	Raw	4	top100	84	236	0	0	0
6	Raw	4	top100	84	236	0	0	0
1	Raw	4	top150	72	200	122	598	4.30116
2	Raw	4	top150	72	200	27	32	0.547723
3	Raw	4	top150	72	200	2	2	0
4	Raw	4	top150	72	200	2	2	0
5	Raw	4	top150	72	200	0	0	0
6	Raw	4	top150	72	200	0	0	0
1	Raw	4	top50	108	308	151	874	4.58258
2	Raw	4	top50	108	308	33	38	0.547723
3	Raw	4	top50	108	308	8	8	0.316228
4	Raw	4	top50	108	308	4	4	0
5	Raw	4	top50	108	308	0	0	0
6	Raw	4	top50	108	308	0	0	0
1	Raw	4	zero	154	883	176	15556	34.3322
2	Raw	4	zero	154	883	173	3832	17.4184
3	Raw	4	zero	154	883	108	434	3.08221
4	Raw	4	zero	154	883	31	40	0.547723
5	Raw	4	zero	154	883	8	8	0
6	Raw	4	zero	154	883	2	2	0
1	Raw	open	arts-prons	221	1031	176	12404	29.577
2	Raw	open	arts-prons	221	1031	166	2330	11.9708
3	Raw	open	arts-prons	221	1031	92	258	2.28035
4	Raw	open	arts-prons	221	1031	19	22	0.316228
5	Raw	open	arts-prons	221	1031	4	4	0
6	Raw	open	arts-prons	221	1031	2	2	0
1	Raw	open	bt	145	422	158	1226	5.95819
2	Raw	open	bt	145	422	40	62	0.774597
3	Raw	open	bt	145	422	15	18	0.316228
4	Raw	open	bt	145	422	4	4	0
5	Raw	open	bt	145	422	0	0	0
6	Raw	open	bt	145	422	0	0	0
1	Raw	open	btb	159	571	172	4618	21.8403
2	Raw	open	btb	159	571	95	674	7.27324
3	Raw	open	btb	159	571	34	48	0.632456
4	Raw	open	btb	159	571	8	8	0
5	Raw	open	btb	159	571	0	0	0
6	Raw	open	btb	159	571	0	0	0
1	Raw	open	top100	147	410	162	1090	5.39444
2	Raw	open	top100	147	410	38	56	0.707107
3	Raw	open	top100	147	410	15	18	0.316228
4	Raw	open	top100	147	410	4	4	0
5	Raw	open	top100	147	410	0	0	0
6	Raw	open	top100	147	410	0	0	0

Link Threshold	Link Type	Span Size	Stopword List	Number of Link Words	Total Links	Number of Bonded Lines	Total Bonds	St Dev
1	Raw	open	top150	128	350	153	902	5.17687
2	Raw	open	top150	128	350	35	52	0.707107
3	Raw	open	top150	128	350	13	16	0.316228
4	Raw	open	top150	128	350	2	2	0
5	Raw	open	top150	128	350	0	0	0
6	Raw	open	top150	128	350	0	0	0
1	Raw	open	top50	188	561	172	1588	6.23699
2	Raw	open	top50	188	561	53	82	0.83666
3	Raw	open	top50	188	561	15	18	0.316228
4	Raw	open	top50	188	561	8	8	0
5	Raw	open	top50	188	561	2	2	0
6	Raw	open	top50	188	561	2	2	0
1	Raw	open	zero	236	1361	176	21602	31.1625
2	Raw	open	zero	236	1361	176	9208	29.3666
3	Raw	open	zero	236	1361	158	2126	11.4848
4	Raw	open	zero	236	1361	108	306	3.16228
5	Raw	open	zero	236	1361	22	28	0.447214
6	Raw	open	zero	236	1361	6	6	0
1	Rel	4	arts-prons	145	584	175	4422	14.3388
2	Rel	4	arts-prons	145	584	107	672	6.44981
3	Rel	4	arts-prons	145	584	35	48	0.632456
4	Rel	4	arts-prons	145	584	4	4	0
5	Rel	4	arts-prons	145	584	4	4	0
6	Rel	4	arts-prons	145	584	4	4	0
1	Rel	4	bt	69	189	122	532	4.06202
2	Rel	4	bt	69	189	19	24	0.447214
3	Rel	4	bt	69	189	4	4	0
4	Rel	4	bt	69	189	2	2	0
5	Rel	4	bt	69	189	0	0	0
6	Rel	4	bt	69	189	0	0	0
1	Rel	4	btb	81	302	154	1942	12.2882
2	Rel	4	btb	81	302	51	480	6.7897
3	Rel	4	btb	81	302	12	12	0.316228
4	Rel	4	btb	81	302	2	2	0
5	Rel	4	btb	81	302	0	0	0
6	Rel	4	btb	81	302	0	0	0
1	Rel	4	top100	71	195	126	572	4.07431
2	Rel	4	top100	71	195	21	26	0.447214
3	Rel	4	top100	71	195	6	6	0
4	Rel	4	top100	71	195	2	2	0
5	Rel	4	top100	71	195	0	0	0
6	Rel	4	top100	71	195	0	0	0
1	Rel	4	top150	60	166	115	484	3.97492
2	Rel	4	top150	60	166	21	26	0.547723
3	Rel	4	top150	60	166	2	2	0
4	Rel	4	top150	60	166	2	2	0
5	Rel	4	top150	60	166	0	0	0
6	Rel	4	top150	60	166	0	0	0
1	Rel	4	top50	93	252	142	668	4.12311
2	Rel	4	top50	93	252	27	32	0.447214
3	Rel	4	top50	93	252	8	8	0.316228

Link Threshold	Link Type	Span Size	Stopword List	Number of Link Words	Total Links	Number of Bonded Lines	Total Bonds	St Dev
4	Rel	4	top50	93	252	4	4	0
5	Rel	4	top50	93	252	0	0	0
6	Rel	4	top50	93	252	0	0	0
1	Rel	4	zero	166	843	176	10732	28.309
2	Rel	4	zero	166	843	162	2166	10.139
3	Rel	4	zero	166	843	90	250	1.84391
4	Rel	4	zero	166	843	29	36	0.447214
5	Rel	4	zero	166	843	8	8	0
6	Rel	4	zero	166	843	4	4	0
1	Rel	open	arts-prons	225	936	176	8164	23.1193
2	Rel	open	arts-prons	225	936	152	1284	8.07465
3	Rel	open	arts-prons	225	936	65	148	1.61245
4	Rel	open	arts-prons	225	936	20	24	0.447214
5	Rel	open	arts-prons	225	936	4	4	0
6	Rel	open	arts-prons	225	936	4	4	0
1	Rel	open	bt	125	339	152	854	4.79583
2	Rel	open	bt	125	339	29	42	0.632456
3	Rel	open	bt	125	339	13	16	0.316228
4	Rel	open	bt	125	339	4	4	0
5	Rel	open	bt	125	339	0	0	0
6	Rel	open	bt	125	339	0	0	0
1	Rel	open	btb	140	488	168	2982	15.2381
2	Rel	open	btb	140	488	77	566	6.67832
3	Rel	open	btb	140	488	28	42	0.632456
4	Rel	open	btb	140	488	10	10	0.316228
5	Rel	open	btb	140	488	0	0	0
6	Rel	open	btb	140	488	0	0	0
1	Rel	open	top100	123	328	154	804	4.58258
2	Rel	open	top100	123	328	30	42	0.632456
3	Rel	open	top100	123	328	13	16	0.316228
4	Rel	open	top100	123	328	4	4	0
5	Rel	open	top100	123	328	0	0	0
6	Rel	open	top100	123	328	0	0	0
1	Rel	open	top150	103	279	144	692	4.40454
2	Rel	open	top150	103	279	28	40	0.632456
3	Rel	open	top150	103	279	11	14	0.316228
4	Rel	open	top150	103	279	2	2	0
5	Rel	open	top150	103	279	0	0	0
6	Rel	open	top150	103	279	0	0	0
1	Rel	open	top50	166	448	164	1064	4.78539
2	Rel	open	top50	166	448	37	52	0.632456
3	Rel	open	top50	166	448	13	16	0.316228
4	Rel	open	top50	166	448	8	8	0
5	Rel	open	top50	166	448	2	2	0
6	Rel	open	top50	166	448	2	2	0
1	Rel	open	zero	251	1306	176	16462	31.1994
2	Rel	open	zero	251	1306	176	5282	20.3199
3	Rel	open	zero	251	1306	142	1056	6.67083
4	Rel	open	zero	251	1306	68	136	1.34164
5	Rel	open	zero	251	1306	26	32	0.447214
6	Rel	open	zero	251	1306	8	8	0

Appendix 4
Concordance Questionnaire

Thanks for taking the time to take part in this survey. I would like you to apply your corpus-linguistic skills to a short analytical task. On the pages which follow are 200 citations of the word 'date', taken from the Bank of English. These have been numbered to aid identification and sorted in several ways to make the task of analysis easier. Common collocates and 'picture' output are also attached. If you would like to use the corpus online to look at this data in other ways (re-sorting, regexp search etc), I am assured that you will get the same set of citations if you use the entire corpus and select 200 lines, but please do not do this yet!

The task that I would like you to perform is as follows:

1) Examine the citations and select the twenty lines which you think are most representative of the behaviour of the node word. Feel free to make use of all the versions of the data. If you are unable to identify twenty lines, select as many as you think are representative.

Please enter the numbers of the citations you select in the boxes below. You do not need to rank the citations, so the order in which you enter them is not important.

2) Re-examine the citations and select twenty which you would be feel would be suitable for use as examples in a dictionary. You may assume that the citations could be edited to some extent, that is, it is possible that only a part of a citation would be used, or that the citation would be expanded to a full sentence. As above, if you feel that there are not twenty useable examples, you may select fewer.

The citations you choose here do not have to overlap with the lines you selected in the previous section, but it does not matter if they do. Please enter the numbers of your selected lines into the boxes below. Here too, order is not important.

3) Finally, please briefly answer the following questions:

i) The data with which you were provided was only a sample of the occurrences of the word 'date', of which there are over 16,000 in total. Do you feel that the sample adequately represented the characteristics of the node word?

You may now check your intuitions against the corpus if you wish. Were they correct?

What size sample would you have chosen for an initial examination of this node word and why would you choose this size?

ii) How many senses of the node word were you able to identify from the 200 citations?

Briefly list which senses you identified.

Do you feel that these are all the senses of this node word?

iii) When you are beginning your analysis of a word (any word), on what grounds (eg total frequency, number of senses expected, distribution across sub-corpora) do you choose the number of sample lines to examine?

If you have any questions relating to this questionnaire, please contact me at: acollier@liverpool.ac.uk.

Thank you once again for your help. I would be grateful if you could return the completed questionnaire to me by Friday 28th July, so that I have enough time to collate the results, which I will present at the Cobuild Seminar on August 9th.

'date' Sample Concordance

001 '' <LTH> Opening up for Metallica on a 65 date US tour, The Cult banged on with their
002 Mead </pres> <prod> Francis Mead </prod> Date Rec: 16 October 1990 Prog No: 90r/32k/0
003 ing details of each match you play. <LTH> DATE COMPETITION OPPONENT VENUE SCORE RESULT
004 ontinue to try to get you into bed. <LTH> Date rape is at the forefront of all our min
005 atre Schedule </h> Playing this weekend:A Date With Judy (1948) Jane Powell plays a vi
006 s had declared their willingness to set a date for starting stage two of economic and
007 he no longer enjoys the preparation for a date. 'Getting ready is part of the fun of i
008 formance Group, who will present We Got A Date, Can't Take Johnny To The Funeral and I
009 July they flew to Fort Worth to perform a date at the Dallas Hilton.Tina was wearing a
010 ate? Almost half thought she should set a date for stepping down; 35 per cent that she
011 and her husband like to sometimes go on a date and spend the night in a hotel.Mrs. Cla
012 a lot of water # At one point he forgot a date that was sort of a simple date on which
013 eggs and flour when he stood her up for a date. But what has been the nature of the fr
014 back to America for the Brando film and a date she wants to keep with Michelle Pfeiffe
015 David Hogan, SM, has been adjourned to a date to be fixed.<dt> 930414 </dt> Cairns at
016 ho has been lined up as his dating agency date unbeknown to wife Alison Steadman who n
017 scope given here is set for this time and date, and for the capital, Paramaribo.<t> <F
018 e. The next regularly slated announcement date is Sept. 18. A brand new directorate wa
019 time. <LTH> JULIAN COPE has added another date to his ''Head On'" tour. It's at Bradfo
020 on, however, with the addition of another date at the London Camden Falcon on Septembe
021 lanation ready for critics of the bizarre date-rape story, 'What Actually Happened # 8
022 g wheels, and as an adult his first blind date.<t> Unidentified Woman (From Radio Ad):
023 oduced banner headlines about his # blind date'' escapade. <t> It was that sense of de
024 HH> LISTINGS <LTH> Concerts are listed by date, then by city # Classical Listings comp
025 ese words and so we will see that sell-by date is no longer associated with perishable
026 n hoarded for decades, only an expert can date a garment. When a skirt length changes,
027 orrect entries selected after the closing date of Tuesday, August 10 will win the new
028 s. The first name drawn after the closing date on October 22, will receive a free Ladd
029 mford Street, London sel 9LS. The closing date is Tuesday August 31, 1993 and the edit
030 es drawn will each receive a kit. Closing date for entries is August 14. Standard rule
031 . Winner to be notified by phone. Closing date? box 29661 <LTH> CZECH 42, passive, liv
032 ied as soon as possible after the closing date. <LTH> 9 Send entries to: AP/Image Hong
033 AB, to reach us no later than the closing date, July 31, 1993. <LTH> 1. TOGECAT <LTH>
034 536 Kings Road, London sw10 OTE. Closing date is August 13. Normal rules apply. <h> P
035 iars Street, London ec88 2NG. <t> Closing date for the contest is January 7, 1992. Sen
036 er pulled from our postbag on the closing date. <t> Over-18's only. News International
037 e than 1500 immediately after the closing date # DAMIEN MARSH .. hard work has paid of
038 s hotel (about $1.2 million # The closing date for inclusion of properties is July 6.B
039 s or call 008 812 772 for details.CLOSING DATE: August 13.DRAWN: August 20.MOBILE PHON
040 the opening rounds. A mutually convenient date should then be set and green fees share
041 FOX> MX.<MOX> Er communications. The copy date for the next issue of Foreword is this
042 :Scotland Yard says the children's deaths date back to 1984. Reports suggest that betw
043 wo years ahead of its intended deployment date. <t> Only ten navigators had been train
044 hy the new Germany has chosen a different date # October 3 # Reunification Day to be a
045 s May 8. You may get a slightly different date by this short cut method than by adding
046 11 begin to accumulate with each dividend date. drps really do serve an important func
047 position figure said the distant election date will give the ruling family time to man
048 ber of parties in parties in the Election date election parliament Albania Mar 23rd 19
049 non-partisan. In announcing the election date, President Roh Tae Woo said there was '
050 peals. <LTH> Make sure you know the final date for accepting a place. Decline unwanted
051 Frank and Barbara Sinatra after the final date of Mr Sinatra's London season. Today he
052 s in Dublin on Saturday should set a firm date for an inter-governmental conference on
053 crowd--don't pick Red Lobster for a first date. Great lunch deals. Hours: Mon.-Thur. a

```
054 <LTH> Do you play it cool after the first date? <LTH> Sarah If it was left that we wou
055 dents had slept with a man on their first date and 39 per cent admitted to being unfai
056 ed agreement with the Chinese on a formal date for the resumption of diplomatic ties,
057  periods of deep loneliness and grief.Her date of birth has been placed somewhere arou
058 gued that it is vital.<p> Dr Salk's ideas date back a long way, but he has linked them
059 ll-wishers, July 26 is the most important date in the year. It marks the anniversary o
060 te, as sensible of the priority of one in date. It was AD 450, that they beat the Scot
061  of [heb.] shows that the poem is late in date. However,Phoenician inscriptions early
062 t, 1960, pp. 181-8).An early inauguration date for the material product concept is ind
063 SENTER: Sue Waldram PRODUCER: Ferri Jahed DATE REC: 10 July 1990 TAPE NO: 90r/32k/028/
064 poser, Sir Andrew Lloyd Webber. <t> Janet Date, a guide and former actress, is in her
065 tempt the same operation again at a later date. He may even, some analysts say,risk th
066 ieved by September 1993 and at some later date the US authorities will declare the sys
067 e people will return to church at a later date.I would like to invite everyone to atte
068 it and its disposition, and of the likely date when the accumulated treasure, with a g
069 e Gulf grows stronger, even if the likely date seems to recede towards the edge of the
070 Franks (S) on design problems (location # date).Contract for Snabl already done, pse i
071 l progressive-punks play a one-off London date with Poisoned Electrick Head at New Cro
072  at 10 cent a share # There is no meeting date set as yet.Pacarc said the issue to WT
073 e has portrayed.'' <t> <h> Benetton's new date;Motor Racing </h> <dt> 25 August 1994 <
074 ork to finish the record, before the next date of the tour in Lisbon. <LTH> The Edge h
075 gn ministry denies there is a hold-up, no date has been set for a new round of talks #
076 ian population is a minority in Serbia.No date has been set for elections and there is
077 ot mention a place. And so far there's no date fixed for the meeting. Until that happe
078  file, no proof, no dossier, no names, no date, no body.And as happens in all hostage
079 y not letting us go to work.<t> Dugan: No date has been set for the resumption of cont
080  by others.<t> 1 Indefinite exclusion: no date is fixed for a return. Consideration of
081 entral government were swiftly put out of date yesterday by the President of Kazakhsta
082 luding them, is discriminating and out of date?They have had some support from leading
083 that the S.E.2000 would be already out of date even before it first flew, and a new de
084  criticisms that their magazine is out of date or has lost its edge # Editor Zanne Zak
085 ] <ZF0> a newspaper it is slightly out of date but erm <F0X> Anything at all it'll giv
086 uly 9) about British tennis are as out of date as the Dunlop Maxply in the attic. <t>
087 r information was always six hours out of date. I get an update from the senior foreca
088 erial. Just as computers overwrite out-of date files on their disks, monks used to scr
089 etting data from instruments years out of date. Small craft allow the use.of up-to-dat
090 hioned administration,traditional, out-of date, a group of elderly men smoking cigars,
091  had a puritan streak, and the concept of date" or 'acquaintance" rape reveals just ho
092 anagers announced the April 5th blast-off date following a flight review at the Kenned
093 in the next three weeks # The NBL cut-off date for the finalisation of imports is next
094 complained that Iraq is offering only one date for a meeting, while he has offered 15
095 al exhaustion of a new mother. An opening date in June would have given her two precio
096 properly revised at the earliest possible date. <LTH> It is also unfortunate that a re
097 this page May 3). Federal credit programs date back to the New Deal, and were meant to
098 > you do you have a sort of a prospective date for having the whole thing up and runni
099 in the Autumn Triangle. <LTH> Provisional date and venue for National Council 1992: 4-
100 FX and FX that therefore the official pub date for the U K <F0X> <ZGY> <F0X> would be
101 res. These measurements also give a quick date for that segment of the whole ice core,
102 o do the 'artwork' for him at the Reading date), going on to waltz until dawn with an
103 ithdrew it and began again, with a record date of Aug. 29. Amdura has challenged the 1
104  shares of Winners on June 20, the record date for Friday's special meeting to conside
105  <ZZ0> Twenty years on from their release date, two albums look set to make this month
106  <ZZ0> <t> Mark Keenan, 28, whose release date had been delayed by 28 days, was found
107 ed in vaults and galleries. Those in Rome date mainly from the third and early fourth
```

sed the 11th hour cancellation of Suede's **date** at the venue, and the closure of the Br
rsonally signed and predated with today's **date**), his eyeglasses, a Koran, a Bible. Fro
at your chance of life was someone else's **date** with </h> death?;Steve Hyett;Part 2 <bl
g article ('Crime made easy') of the same **date** seems to have that problem. Can you exp
ill follow their previously announced six **date** tour, are priced <KPD> 8.50. Fans will
proposed October the 30th as the starting **date** # Mrs # Mandela's lawyer argued success
e of a visit with my son on such and such **date** else I would have been there. Probation
e Nile, the village is surrounded by tall **date** palms and lush green farmland. Its narr
reed last night to set 2000 as the target **date** for stabilising emissions of carbon dio
ng Board and set an implementation target **date** of January 1. <t> The working party hop
Evans has two weeks from his termination **date** to appeal the decision. As for Randall,
the hatching of meadow birds. After that **date** the mechanical cultivation of fields wi
, however, always say that he shares that **date** with my wife. <LTH> Dr L Keith France <
h the Council by 1 April 1993. After that **date**, it will be an offence to run an unregi
er the first use of their cards from that **date**. They have until March 1, 1993, to clea
decided Tuesday to have a meeting on that **date**, the judge ordered the meeting held and
e lever. <CQ0> <t> Leap-horn gave him the **date** of the death of Pointed Shoes.<t> <CQ1>
Ted was a model patient, remembering the **date** of every appointment, and following a lo
1;one from the trial judge announcing the **date** of his execution in six weeks and one f
ld is <CES> <t> So it would seem from the **date** of his birth # <t> My God # <t> He's no
of work on a regular basis and is not the **date** his family or household goods and effec
terward, Leo asked her for a date,and the **date** led to this. This deal has to be cash,
that Tunisia has implicitly accepted the **date** of the 27th, so that would suggest that
d, accompanied by a printed report of the **date**, time and number of the attempted conne
hecking my face again when he came to the **date** of birth, turning to the back to see th
ers, like the driver of a Hansom cab. The **date** is 29 March 1920.<t> <FCH> Above left <
rsvermehrung beim Umbau," which bears the **date** December 13, 1932 at the end.39 Ibid.40
than the exact calendar months after the **date** the loan was opened. <LTH> Written quot
if you post your order and payment by the **date** on the enclosed form. <LTH> <FCH> But w
dvert and your order # with a note of the **date** you sent it. Don't forget to give your
ably about 13 # I can't remember what the **date** on that is--about 1773 or so # She--she
round war against Iraq # He also said the **date** is imminent and that a ground war can't
ls in the--on the--petrified tree and the **date** # And in all of my trips out to Montana
tain intervention until September 20, the **date** of the referendum, if necessary, and wh
Mrs Thatcher acted, bringing forward the **date** for a possible leadership election in o
jor is expected to confirm April 9 as the **date** of the election. <h> Tories pin electio
will debate the issue.<t> But setting the **date** is seen as little more than a palliativ
rybody else appears to have forgotten the **date**. Others feel the need to discuss the em
nt Assembly. The elections, such as their **date** and the voters' roll and even that meet
know. Leah claimed she knew by the third **date** that she wanted to become Mrs Winter. B
s, he asks you out again despite no third **date** action), you know you've built a founda
39 (London,1981). Berlitz associated this **date** with the dire predictions given in Grib
ndependence, and Chart 91 is set for this **date** for Helsinki, the capital, for 12.00 no
kend Edition # I'm Neal Conan.<t> On this **date** in 1956 the Republican Party nominated
ecretary of state, brought Franklin up to **date** on the bloodshed in his beloved France.
hout adequate nuclear weapons, kept up to **date** and based forward in Europe, our defenc
and he said he would bring Mr Bush up to **date** on the issue:If we were forced to resor
o you work in the city?" and so on.<t> To **date** no machine has successfully fooled an e
eave to enable the returner to keep up to **date** with developments.<t> Various other sch
thought of her man # <SO> Very much up to **date**, only been in service with our own forc
miles with all the service records up to **date**. Abandoned only because arthritis had g
cers so they can keep their members up to **date** with what is happening in the industry.
contributed just less than dollars 3m to **date**. Most big state campaigns cost about do
Boy In New Cross # their greatest hit to **date**, is nowhere to be seen; but they do squ

```
162 th their good work. Their achievements to date are quite amazing: land reclaimed; gard
163  to teach has not been very productive to date, nor is it likely to become more so in
164 anwhile, the multi-national menu is up-to date without being trendy: strikingly fresh
165  <LTH> Sir-We are almost there, having to date raised <KPD> 92,000 in aid of lifeboat
166 most notable two year-old performances to date # writes Dean Bailey <LTH> RESPONDING g
167  three billion bases, or coding units. To date, fossilised DNA has been extracted from
168 t to be different. <LTH> As we come up to date, people # do it'' to be the same. Long
169 ting financial climate in the country. To date I have only received five applications
170 rame the most impassioned Vedder vocal to date. He creates an opening mood of loneline
171 ficant clubs, this is Brainiak's story to date # B Sides'' features three God-blessed
172 at' rhythm, 'Love # is his biggest hit to date. Chang's brand of lyrics label him as a
173 ng really because er it's just been up-to date and it <M01> Mhm.<M02> I mean that's it
174 In India where we've got reasonably up to date statistics on population. Er there's be
175 tivities in which you've been involved to date?<M01> Er the spectrum of that would rea
176  are now available but ask somebody up-to date.<M01> Mm.<F01> And of course computeriz
177 utlining my 'firm grasp of the most up-to date trauma procedures". <t> The references
178  and we are encouraged by our progress to date." In New York, John S. Reidy, analyst f
179 agements. In one of the few such moves to date, KGF recently moved the management of B
180 # Mrs Milosevic's only live appearance to date came on one of its interview shows # an
181 e full potential has not been realised to date owing to the ground.<t> In the Temple S
182 decriminalising of breaches in the law to date.<t> The ruling Christian Democrats and
183 ov has won almost all their encounters to date.<t> Short emerged from the candidates'
184  even though the home loan was paid up to date. Few would ever have imagined they coul
185  possible # says Bremner. The evidence to date, he says,suggests that men given a comp
186 ery 3 months; none has become infected to date. In more than 70 incidents worldwide in
187 verns, and to considerable depths, but to date no detailed studies have been made of t
188 world, and though this is the first up-to date survey of its politics, it does not pro
189 at makes it the No. 1 film of the year to date and the biggest April release in the hi
190 ei where tier upon tier of rock-cut tombs date from as long ago as the thirteenth cent
191 mount of ownership.<t> Levinson: No trial date has been set yet for the Janis lawsuit
192 eaplane 2222 No built <FCH> Aircraft Type Date Purpose of Design No built <FCH> Anson
193 meet the announced 20 March maiden voyage date. On 10 October the company released a n
194 ents occur in the group of kouroi that we date the earliest.<t> JAFFE: The torso, on t
195 er-finals in Filderstadt. <t> <h> Wembley date;Rugby League </h> <dt> 15 October 1994
196  trying to block a money-spinning Wembley date.Edwards hopes to convince FA Cup semi-f
197 rs # It's not stated clearly back to what date this is effective # The decrees come on
198 and we started restoring our murals which date back to the Portuguese era in the 14th
199 ember after ordering the reactors # which date back to the 1950s # to be shut. They po
200 h, blossomed in the presence of women who date act ors and princes, dine in Milan and
```

Appendix 5a
Pearson's Correlation Coefficient (r) for Automatic vs Manual Analyses

Pearson's Correlation Coefficient (r) for Automatic vs Manual Analyses

Stopword List	Link Threshold	Link Type	Compared with	r
zero	4	Raw	Repr	0.160472
zero	5	Raw	Repr	0.158715
arts-prons	3	Raw	Usable	0.151627
zero	3	Raw	Repr	0.140843
arts-prons	4	Raw	Usable	0.139588
arts-prons	2	Raw	Usable	0.137518
zero	2	Raw	Repr	0.137459
zero	3	Abs	Repr	0.134821
zero	3	Abs	Usable	0.133738
btb	2	Rel	Usable	0.13083
zero	5	Raw	Usable	0.12586
zero	4	Raw	Usable	0.125596
btb	2	Raw	Usable	0.124091
arts-prons	2	Raw	Repr	0.120067
arts-prons	1	Raw	Repr	0.119728
arts-prons	2	Abs	Usable	0.11948
top50	1	Rel	Usable	0.11899
zero	6	Raw	Usable	0.118291
top50	1	Raw	Usable	0.116674
btb	1	Abs	Usable	0.116009
btb	1	Abs	Repr	0.115263
arts-prons	1	Raw	Usable	0.113599
arts-prons	2	Rel	Usable	0.113581
bt	1	Raw	Repr	0.112679
arts-prons	4	Raw	Repr	0.112003
zero	6	Raw	Repr	0.111595
top50	1	Abs	Usable	0.107549
zero	1	Raw	Repr	0.106711
btb	2	Rel	Repr	0.103401
arts-prons	3	Raw	Repr	0.102863
btb	2	Abs	Usable	0.101283
btb	3	Raw	Usable	0.100125
btb	2	Raw	Repr	0.0952839
zero	2	Abs	Repr	0.0937013
zero	3	Raw	Usable	0.0933104
zero	2	Abs	Usable	0.0920593
zero	2	Raw	Usable	0.0919059
bt	5	Raw	Repr	0.0914492
btb	5	Raw	Repr	0.0914492
top50	6	Raw	Repr	0.0914492
top50	2	Raw	Repr	0.0900972
arts-prons	1	Rel	Usable	0.089265
btb	2	Abs	Repr	0.0863372
arts-prons	1	Abs	Usable	0.0862142
arts-prons	3	Rel	Usable	0.0860917
btb	3	Raw	Repr	0.0848565
zero	1	Raw	Usable	0.084021

Pearson's Correlation Coefficient (r) for Automatic vs Manual Analyses

Stopword List	Link Threshold	Link Type	Compared with	r
arts-prons	2	Abs	Repr	0.0810827
bt	1	Rel	Repr	0.0809557
bt	1	Raw	Usable	0.0807244
top50	1	Abs	Repr	0.0773212
zero	4	Abs	Usable	0.0772738
top150	1	Raw	Repr	0.0762544
top50	5	Raw	Repr	0.0743581
zero	4	Rel	Repr	0.0743539
top100	1	Raw	Repr	0.0741777
bt	4	Raw	Repr	0.0737642
top50	1	Raw	Repr	0.0721254
top50	2	Raw	Usable	0.0717595
top50	1	Rel	Repr	0.0717154
bt	1	Abs	Usable	0.0698631
top100	2	Raw	Usable	0.0684238
zero	1	Rel	Repr	0.0669735
arts-prons	6	Raw	Usable	0.0663645
bt	5	Raw	Usable	0.0656034
btb	5	Raw	Usable	0.0656034
top50	6	Raw	Usable	0.0656034
top100	1	Raw	Usable	0.0642593
btb	1	Raw	Usable	0.0634492
arts-prons	1	Rel	Repr	0.0612989
top150	2	Raw	Usable	0.0608767
zero	2	Rel	Repr	0.0599399
bt	4	Raw	Usable	0.0597867
bt	1	Rel	Usable	0.0595384
btb	4	Raw	Repr	0.0580086
btb	4	Raw	Usable	0.0575427
zero	3	Rel	Repr	0.057217
top150	1	Raw	Usable	0.05598
top50	4	Raw	Usable	0.0559443
top50	2	Abs	Usable	0.055499
arts-prons	2	Rel	Repr	0.0552844
btb	3	Rel	Repr	0.0548648
bt	2	Raw	Usable	0.0547751
btb	1	Rel	Usable	0.0532424
arts-prons	4	Abs	Usable	0.0520662
arts-prons	5	Abs	Usable	0.0520662
top50	2	Rel	Repr	0.0517323
zero	4	Rel	Usable	0.050919
btb	3	Rel	Usable	0.0469044
top100	4	Raw	Usable	0.0460898
top150	4	Raw	Usable	0.0460898
top100	2	Rel	Repr	0.0460015
arts-prons	5	Rel	Usable	0.0457302
arts-prons	5	Raw	Usable	0.0453004
top100	2	Raw	Repr	0.0451467
zero	1	Rel	Usable	0.0443776
top100	2	Rel	Usable	0.0427069

Pearson's Correlation Coefficient (r) for Automatic vs Manual Analyses

Stopword List	Link Threshold	Link Type	Compared with	r
arts-prons	6	Rel	Usable	0.041473
zero	5	Abs	Usable	0.0387049
top50	4	Raw	Repr	0.038473
arts-prons	6	Raw	Repr	0.0380452
top100	4	Raw	Repr	0.0379376
top150	4	Raw	Repr	0.0379376
top150	2	Raw	Repr	0.0376191
top50	5	Raw	Usable	0.0358538
arts-prons	1	Abs	Repr	0.0355865
top150	2	Rel	Repr	0.0348022
bt	3	Rel	Repr	0.0323498
top150	2	Rel	Usable	0.0314265
bt	3	Raw	Repr	0.031341
top50	3	Rel	Repr	0.0308943
arts-prons	3	Abs	Usable	0.0295721
btb	1	Raw	Repr	0.0293713
zero	2	Rel	Usable	0.0288602
bt	2	Abs	Repr	0.0285911
arts-prons	5	Raw	Repr	0.0284467
top50	2	Rel	Usable	0.0269155
top50	3	Rel	Usable	0.0267172
bt	1	Abs	Repr	0.0266652
top100	1	Rel	Usable	0.0260982
zero	3	Rel	Usable	0.0254732
bt	2	Raw	Repr	0.0247631
zero	5	Rel	Repr	0.0238316
zero	5	Rel	Usable	0.0227609
arts-prons	3	Rel	Repr	0.0206558
top100	3	Raw	Repr	0.0200875
top150	3	Raw	Repr	0.0200875
top100	2	Abs	Repr	0.0190385
top150	2	Abs	Repr	0.0190385
bt	2	Abs	Usable	0.0186653
top150	1	Abs	Usable	0.0185879
bt	3	Raw	Usable	0.0176936
top100	1	Abs	Usable	0.0168762
top100	1	Rel	Repr	0.0152197
zero	1	Abs	Usable	0.0152132
zero	4	Abs	Repr	0.0125578
top100	3	Raw	Usable	0.012112
top150	3	Raw	Usable	0.012112
btb	1	Rel	Repr	0.0116549
arts-prons	4	Rel	Usable	0.010805
top50	4	Rel	Repr	0.0089975
top150	1	Rel	Repr	0.00780905
bt	3	Rel	Usable	0.00610162
top100	2	Abs	Usable	0.00586106
top150	2	Abs	Usable	0.00586106
top50	4	Rel	Usable	0.00573766
bt	4	Rel	Usable	0.00494328

Pearson's Correlation Coefficient (r) for Automatic vs Manual Analyses

Stopword List	Link Threshold	Link Type	Compared with	r
top100	4	Rel	Usable	0.00494328
top150	4	Rel	Usable	0.00494328
top50	2	Abs	Repr	0.00491112
top150	1	Rel	Usable	0.00410956
top50	3	Raw	Usable	0.0009486
zero	1	Abs	Repr	0.000846477
zero	6	Rel	Repr	0.000771992
bt	6	Raw	Repr	0.0
bt	6	Raw	Usable	0.0
bt	6	Abs	Repr	0.0
bt	6	Abs	Usable	0.0
bt	6	Rel	Repr	0.0
bt	6	Rel	Usable	0.0
btb	6	Raw	Repr	0.0
btb	6	Raw	Usable	0.0
btb	6	Abs	Repr	0.0
btb	6	Abs	Usable	0.0
btb	6	Rel	Repr	0.0
btb	6	Rel	Usable	0.0
top100	5	Raw	Repr	0.0
top100	5	Raw	Usable	0.0
top100	5	Abs	Repr	0.0
top100	5	Abs	Usable	0.0
top100	5	Rel	Repr	0.0
top100	5	Rel	Usable	0.0
top100	6	Raw	Repr	0.0
top100	6	Raw	Usable	0.0
top100	6	Abs	Repr	0.0
top100	6	Abs	Usable	0.0
top100	6	Rel	Repr	0.0
top100	6	Rel	Usable	0.0
top150	5	Raw	Repr	0.0
top150	5	Raw	Usable	0.0
top150	5	Abs	Repr	0.0
top150	5	Abs	Usable	0.0
top150	5	Rel	Repr	0.0
top150	5	Rel	Usable	0.0
top150	6	Raw	Repr	0.0
top150	6	Raw	Usable	0.0
top150	6	Abs	Repr	0.0
top150	6	Abs	Usable	0.0
top150	6	Rel	Repr	0.0
top150	6	Rel	Usable	0.0
top50	5	Abs	Repr	0.0
top50	5	Abs	Usable	0.0
top50	5	Rel	Repr	0.0
top50	5	Rel	Usable	0.0
top50	6	Abs	Repr	0.0
top50	6	Abs	Usable	0.0
top50	6	Rel	Repr	0.0

Pearson's Correlation Coefficient (r) for Automatic vs Manual Analyses

Stopword List	Link Threshold	Link Type	Compared with	r
top50	6	Rel	Usable	0.0
bt	2	Rel	Usable	-0.000902095
bt	2	Rel	Repr	-0.00149611
zero	6	Rel	Usable	-0.00276863
arts-prons	3	Abs	Repr	-0.00469006
arts-prons	4	Abs	Repr	-0.00529693
arts-prons	5	Abs	Repr	-0.00529693
bt	3	Abs	Repr	-0.0060907
bt	4	Rel	Repr	-0.0060907
btb	3	Abs	Repr	-0.0060907
top100	3	Abs	Repr	-0.0060907
top100	4	Rel	Repr	-0.0060907
top150	3	Abs	Repr	-0.0060907
top150	4	Rel	Repr	-0.0060907
top50	3	Abs	Repr	-0.0060907
zero	5	Abs	Repr	-0.0062203
arts-prons	6	Rel	Repr	-0.00666516
top100	3	Rel	Repr	-0.00722152
top150	3	Rel	Repr	-0.00722152
arts-prons	5	Rel	Repr	-0.00798018
top150	1	Abs	Repr	-0.0102665
top50	3	Raw	Repr	-0.0181693
top100	1	Abs	Repr	-0.0183122
top100	3	Rel	Usable	-0.0201881
top150	3	Rel	Usable	-0.0201881
btb	4	Rel	Usable	-0.0261382
arts-prons	4	Rel	Repr	-0.0270862
bt	3	Abs	Usable	-0.0316736
btb	3	Abs	Usable	-0.0316736
top100	3	Abs	Usable	-0.0316736
top150	3	Abs	Usable	-0.0316736
top50	3	Abs	Usable	-0.0316736
btb	4	Rel	Repr	-0.0392034
arts-prons	6	Abs	Usable	-0.0599534
bt	4	Abs	Usable	-0.0599534
bt	5	Abs	Usable	-0.0599534
bt	5	Rel	Usable	-0.0599534
btb	4	Abs	Usable	-0.0599534
btb	5	Abs	Usable	-0.0599534
btb	5	Rel	Usable	-0.0599534
top100	4	Abs	Usable	-0.0599534
top150	4	Abs	Usable	-0.0599534
top50	4	Abs	Usable	-0.0599534
zero	6	Abs	Usable	-0.0599534
arts-prons	6	Abs	Repr	-0.0667674
bt	4	Abs	Repr	-0.0667674
bt	5	Abs	Repr	-0.0667674
bt	5	Rel	Repr	-0.0667674
btb	4	Abs	Repr	-0.0667674
btb	5	Abs	Repr	-0.0667674

Pearson's Correlation Coefficient (r) for Automatic vs Manual Analyses

Stopword List	Link Threshold	Link Type	Compared with	r
btb	5	Rel	Repr	-0.0667674
top100	4	Abs	Repr	-0.0667674
top150	4	Abs	Repr	-0.0667674
top50	4	Abs	Repr	-0.0667674
zero	6	Abs	Repr	-0.0667674

Appendix 5b
Simple Rank Scores based on Top 10 Representative Lines from 'date' Sample

Simple Rank Scores based on Top 10 Representative Lines from 'date' Sample

Stopword List	No. of Links	Span Type	Link Type	Score
arts-prons	2	open	abs	0.786701
arts-prons	4	open	raw	0.781586
arts-prons	3	open	raw	0.781074
arts-prons	2	open	raw	0.776471
arts-prons	3	open	abs	0.765729
arts-prons	2	open	rel	0.764706
arts-prons	3	open	rel	0.760614
top50	1	open	abs	0.757033
arts-prons	1	open	raw	0.755499
bt	1	open	abs	0.748849
arts-prons	5	open	raw	0.742199
arts-prons	1	open	rel	0.741176
arts-prons	4	open	rel	0.73913
top150	1	open	abs	0.729412
top100	1	open	abs	0.714066
zero	3	open	abs	0.698721
top50	3	open	raw	0.686957
arts-prons	1	open	abs	0.680818
zero	2	open	abs	0.670588
zero	5	open	raw	0.667008
btb	2	open	abs	0.662916
btb	3	open	raw	0.66087
bt	2	open	rel	0.659847
top50	1	open	rel	0.658824
arts-prons	5	open	rel	0.65422
bt	1	open	rel	0.652685
zero	4	open	raw	0.652174
zero	3	open	raw	0.650128
btb	1	open	abs	0.648593
zero	6	open	raw	0.648082
btb	3	open	rel	0.646547
zero	4	open	abs	0.636829
top50	2	open	raw	0.632737
arts-prons	1	fixed	abs	0.631714
arts-prons	1	fixed	raw	0.631714
arts-prons	1	fixed	rel	0.631714
arts-prons	2	fixed	abs	0.631714
arts-prons	2	fixed	raw	0.631714
arts-prons	2	fixed	rel	0.631714
arts-prons	3	fixed	raw	0.631714
bt	1	fixed	abs	0.631714
bt	1	fixed	raw	0.631714
bt	1	fixed	rel	0.631714
btb	1	fixed	abs	0.631714
btb	1	fixed	raw	0.631714
btb	1	fixed	rel	0.631714
top100	1	fixed	abs	0.631714
top100	1	fixed	raw	0.631714

Simple Rank Scores based on Top 10 Representative Lines from 'date' Sample

Stopword List	No. of Links	Span Type	Link Type	Score
top100	1	fixed	rel	0.631714
top150	1	fixed	abs	0.631714
top150	1	fixed	raw	0.631714
top150	1	fixed	rel	0.631714
top50	1	fixed	abs	0.631714
top50	1	fixed	raw	0.631714
top50	1	fixed	rel	0.631714
zero	1	fixed	abs	0.631714
zero	1	fixed	raw	0.631714
zero	1	fixed	rel	0.631714
zero	2	fixed	abs	0.631714
zero	2	fixed	raw	0.631714
zero	2	fixed	rel	0.631714
zero	3	fixed	abs	0.631714
zero	3	fixed	raw	0.631714
zero	3	fixed	rel	0.631714
zero	2	open	raw	0.630691
top50	2	open	rel	0.626598
bt	3	open	raw	0.625064
btb	2	open	raw	0.623529
top50	1	open	raw	0.618926
btb	2	open	rel	0.615345
bt	2	open	raw	0.613811
top100	2	open	rel	0.605627
zero	5	open	rel	0.605115
zero	1	open	raw	0.603069
zero	4	open	rel	0.602046
top50	2	open	abs	0.6
bt	1	open	raw	0.597954
zero	3	open	rel	0.595396
top50	3	open	rel	0.594373
top100	1	open	rel	0.58977
top150	1	open	rel	0.589258
zero	1	open	rel	0.585166
zero	2	open	rel	0.582609
top150	2	open	rel	0.578005
btb	1	open	rel	0.571867
btb	1	open	raw	0.569821
btb	4	open	raw	0.569821
top100	2	open	raw	0.568286
arts-prons	4	open	abs	0.555499
zero	1	open	abs	0.551407
arts-prons	6	open	raw	0.550384
zero	6	open	rel	0.549872
top50	4	open	raw	0.543223
top50	2	fixed	raw	0.541688
top50	2	fixed	rel	0.541688
top100	3	open	raw	0.536573
arts-prons	3	fixed	rel	0.530946
zero	4	fixed	raw	0.530946
btb	3	open	abs	0.517136

Simple Rank Scores based on Top 10 Representative Lines from 'date' Sample

Stopword List	No. of Links	Span Type	Link Type	Score
top150	2	open	raw	0.511509
top100	1	open	raw	0.502302
btb	4	open	rel	0.498721
arts-prons	6	open	rel	0.491049
zero	4	fixed	rel	0.472634
top150	1	open	raw	0.470077
btb	2	fixed	abs	0.468031
btb	2	fixed	raw	0.468031
btb	2	fixed	rel	0.468031
top50	2	fixed	abs	0.468031
bt	3	open	rel	0.466496
arts-prons	3	fixed	abs	0.457289
bt	2	open	abs	0.437852
top100	2	open	abs	0.414834
top150	3	open	raw	0.374936
bt	2	fixed	raw	0.371867
zero	5	open	abs	0.363683
top150	2	open	abs	0.345269
top100	3	open	rel	0.344246
btb	5	open	raw	0.341688
top50	4	open	rel	0.340153
arts-prons	4	fixed	raw	0.329412
btb	3	fixed	raw	0.313555
zero	4	fixed	abs	0.308951
arts-prons	4	fixed	rel	0.299744
top100	2	fixed	raw	0.287468
top100	2	fixed	rel	0.287468
zero	6	open	abs	0.286957
arts-prons	5	open	abs	0.280818
top100	2	fixed	abs	0.273657
bt	2	fixed	rel	0.2711
top150	3	open	rel	0.258824
arts-prons	4	fixed	abs	0.209719
top150	2	fixed	abs	0.204092
top150	2	fixed	raw	0.204092
top150	2	fixed	rel	0.204092
bt	2	fixed	abs	0.198977
top150	4	open	raw	0.194373
top100	4	open	raw	0.193862
arts-prons	6	open	abs	0.193862
top150	6	open	raw	0.192839
top100	6	open	raw	0.192327
top150	4	open	rel	0.191816
bt	5	open	raw	0.190793
top150	5	open	raw	0.190793
top50	6	open	raw	0.190793
top100	4	open	rel	0.18977
top150	5	open	rel	0.18977
top100	5	open	raw	0.188747
top100	5	open	rel	0.188235
top50	5	open	rel	0.186701

Simple Rank Scores based on Top 10 Representative Lines from 'date' Sample

Stopword List	No. of Links	Span Type	Link Type	Score
bt	4	open	rel	0.184655
top50	5	open	raw	0.184143
btb	3	fixed	abs	0.18312
btb	3	fixed	rel	0.18312
bt	4	open	raw	0.181074
top50	3	open	abs	0.175448
zero	5	fixed	abs	0.150895
zero	5	fixed	raw	0.150895
zero	5	fixed	rel	0.150895
top50	3	fixed	raw	0.143223
bt	3	fixed	raw	0.129412
top100	3	fixed	raw	0.129412
top150	3	fixed	raw	0.129412
arts-prons	5	fixed	abs	0.121228
arts-prons	5	fixed	raw	0.121228
arts-prons	5	fixed	rel	0.121228
zero	6	fixed	abs	0.121228
zero	6	fixed	raw	0.121228
zero	6	fixed	rel	0.121228
top50	3	fixed	rel	0.113555
bt	6	open	abs	0.100256
bt	6	open	raw	0.0997442
btb	6	open	raw	0.0997442
bt	3	fixed	abs	0.0997442
bt	3	fixed	rel	0.0997442
bt	6	open	rel	0.0997442
btb	4	fixed	abs	0.0997442
btb	4	fixed	raw	0.0997442
btb	4	fixed	rel	0.0997442
btb	6	open	abs	0.0997442
top100	3	fixed	abs	0.0997442
top100	3	fixed	rel	0.0997442
top150	3	fixed	abs	0.0997442
top150	3	fixed	rel	0.0997442
top50	3	fixed	abs	0.0997442
btb	5	open	rel	0.0992327
btb	6	open	rel	0.0992327
top50	6	open	abs	0.0992327
top50	6	open	rel	0.0992327
bt	5	open	rel	0.0987212
top100	5	open	abs	0.0987212
top150	5	open	abs	0.0987212
top50	5	open	abs	0.0987212
bt	5	open	abs	0.0982097
bt	4	open	abs	0.0976982
btb	5	open	abs	0.0971867
top100	4	open	abs	0.0971867
top150	4	open	abs	0.0971867
top50	4	open	abs	0.0971867
top100	6	open	rel	0.0966752
top150	6	open	rel	0.0966752

Simple Rank Scores based on Top 10 Representative Lines from 'date' Sample

Stopword List	No. of Links	Span Type	Link Type	Score
bt	3	open	abs	0.0961637
btb	4	open	abs	0.0961637
top150	3	open	abs	0.0961637
top100	3	open	abs	0.0951407
top150	6	open	abs	0.0936061
top100	6	open	abs	0.0925831
arts-prons	6	fixed	abs	0.0214834
arts-prons	6	fixed	raw	0.0214834
arts-prons	6	fixed	rel	0.0214834
bt	4	fixed	abs	0
bt	4	fixed	raw	0
bt	4	fixed	rel	0
bt	5	fixed	abs	0
bt	5	fixed	raw	0
bt	5	fixed	rel	0
bt	6	fixed	abs	0
bt	6	fixed	raw	0
bt	6	fixed	rel	0
btb	5	fixed	abs	0
btb	5	fixed	raw	0
btb	5	fixed	rel	0
btb	6	fixed	abs	0
btb	6	fixed	raw	0
btb	6	fixed	rel	0
top100	4	fixed	abs	0
top100	4	fixed	raw	0
top100	4	fixed	rel	0
top100	5	fixed	abs	0
top100	5	fixed	raw	0
top100	5	fixed	rel	0
top100	6	fixed	abs	0
top100	6	fixed	raw	0
top100	6	fixed	rel	0
top150	4	fixed	abs	0
top150	4	fixed	raw	0
top150	4	fixed	rel	0
top150	5	fixed	abs	0
top150	5	fixed	raw	0
top150	5	fixed	rel	0
top150	6	fixed	abs	0
top150	6	fixed	raw	0
top150	6	fixed	rel	0
top50	4	fixed	abs	0
top50	4	fixed	raw	0
top50	4	fixed	rel	0
top50	5	fixed	abs	0
top50	5	fixed	raw	0
top50	5	fixed	rel	0
top50	6	fixed	abs	0
top50	6	fixed	raw	0
top50	6	fixed	rel	0

Appendix 5c
Simple Rank Scores based on Top 9 Usable Lines from 'date' Sample

Simple Rank Scores based on Top 9 Usable Lines from 'date' Sample

Stopword List	No. of Links	Span Type	Link Type	Score
arts-prons	2	fixed	abs	0.818594
arts-prons	2	open	abs	0.81746
arts-prons	2	fixed	rel	0.786281
arts-prons	3	fixed	raw	0.755102
top50	1	fixed	abs	0.752834
top50	1	fixed	raw	0.751134
arts-prons	3	open	abs	0.750567
arts-prons	2	fixed	raw	0.744898
top50	1	fixed	rel	0.736961
top50	1	open	abs	0.735261
zero	2	fixed	abs	0.727891
arts-prons	3	fixed	rel	0.717687
arts-prons	3	open	raw	0.716553
arts-prons	4	open	raw	0.71542
arts-prons	2	open	raw	0.705782
arts-prons	3	open	rel	0.705782
arts-prons	2	open	rel	0.704649
zero	3	fixed	abs	0.696712
arts-prons	1	open	abs	0.689909
arts-prons	4	open	rel	0.685941
top50	1	open	rel	0.684807
arts-prons	1	open	raw	0.679138
arts-prons	1	fixed	abs	0.676304
arts-prons	1	open	rel	0.671769
zero	2	open	abs	0.669501
top50	3	open	raw	0.666667
zero	3	open	abs	0.664966
zero	3	fixed	rel	0.661565
top50	2	fixed	raw	0.656463
arts-prons	5	open	raw	0.654195
btb	1	open	abs	0.645692
arts-prons	3	fixed	abs	0.642857
btb	1	fixed	abs	0.64229
bt	1	fixed	abs	0.629819
top50	2	open	rel	0.628118
top50	2	open	raw	0.62415
top50	1	open	raw	0.622449
arts-prons	1	fixed	rel	0.618481
top50	2	fixed	rel	0.613379
arts-prons	1	fixed	raw	0.608277
bt	1	open	abs	0.60034
btb	2	open	abs	0.599206
bt	1	fixed	rel	0.594671
btb	3	open	raw	0.589569
bt	1	fixed	raw	0.584467
btb	3	open	rel	0.580499
top50	2	open	abs	0.580499
btb	1	fixed	rel	0.579365

Simple Rank Scores based on Top 9 Usable Lines from 'date' Sample

Stopword List	No. of Links	Span Type	Link Type	Score
top50	2	fixed	abs	0.579365
btb	2	fixed	abs	0.578231
zero	4	open	abs	0.571429
btb	1	open	rel	0.570862
btb	2	open	raw	0.570295
btb	1	open	raw	0.566893
btb	2	open	rel	0.566327
top50	3	open	rel	0.565193
zero	1	fixed	abs	0.565193
top100	1	open	abs	0.564626
zero	3	fixed	raw	0.562358
zero	1	open	abs	0.558957
btb	2	fixed	rel	0.557823
zero	4	fixed	raw	0.556122
top100	3	open	raw	0.552721
btb	1	fixed	raw	0.55102
zero	2	fixed	rel	0.55102
top150	1	open	abs	0.549887
top150	1	fixed	abs	0.548753
btb	2	fixed	raw	0.546485
arts-prons	5	open	rel	0.545351
zero	5	open	raw	0.543084
top100	2	open	rel	0.539683
top100	1	open	rel	0.537982
top100	1	fixed	abs	0.536848
zero	4	open	raw	0.534014
zero	4	fixed	rel	0.533447
bt	2	open	rel	0.532313
zero	3	open	raw	0.532313
arts-prons	4	open	abs	0.531179
bt	1	open	rel	0.530045
zero	4	open	rel	0.524376
zero	5	open	rel	0.52381
top150	2	open	rel	0.518707
zero	2	open	raw	0.518707
zero	3	open	rel	0.518707
top100	2	open	raw	0.514739
zero	4	fixed	abs	0.513605
btb	4	open	raw	0.512472
zero	2	open	rel	0.505669
top150	1	open	rel	0.502268
zero	1	open	rel	0.501701
zero	6	open	raw	0.497166
bt	3	open	raw	0.496599
zero	1	open	raw	0.496599
top100	1	fixed	raw	0.480726
bt	2	open	raw	0.479025
top100	1	fixed	rel	0.473356
arts-prons	4	fixed	raw	0.467687
btb	3	open	abs	0.466553
top150	1	fixed	raw	0.466553

Simple Rank Scores based on Top 9 Usable Lines from 'date' Sample

Stopword List	No. of Links	Span Type	Link Type	Score
bt	1	open	raw	0.459184
top150	1	fixed	rel	0.450113
top150	2	open	raw	0.446712
top100	1	open	raw	0.442744
zero	2	fixed	raw	0.429138
zero	6	open	rel	0.418367
arts-prons	6	open	raw	0.410998
top50	4	open	raw	0.406463
zero	1	fixed	rel	0.389456
top150	1	open	raw	0.379819
arts-prons	4	fixed	abs	0.379252
top100	2	fixed	raw	0.377551
arts-prons	4	fixed	rel	0.376417
top100	2	fixed	rel	0.372449
btb	4	open	rel	0.36678
bt	2	fixed	raw	0.365646
bt	2	fixed	rel	0.352608
zero	1	fixed	raw	0.333333
top150	3	open	raw	0.30839
btb	3	fixed	raw	0.29932
zero	5	open	abs	0.298186
top100	2	open	abs	0.292517
top100	2	fixed	abs	0.284014
top50	3	fixed	raw	0.281746
top150	2	open	abs	0.280612
arts-prons	6	open	rel	0.280045
top100	3	open	rel	0.274376
bt	3	open	rel	0.271542
btb	5	open	raw	0.269274
bt	2	open	abs	0.251134
zero	5	fixed	abs	0.211451
zero	6	open	abs	0.211451
zero	5	fixed	rel	0.210884
arts-prons	5	open	abs	0.206916
btb	3	fixed	abs	0.206916
zero	5	fixed	raw	0.204082
btb	3	fixed	rel	0.201247
top150	2	fixed	raw	0.200113
top150	2	fixed	abs	0.190476
top150	2	fixed	rel	0.189909
bt	2	fixed	abs	0.185374
top50	4	open	rel	0.18424
top150	3	open	rel	0.179705
arts-prons	6	fixed	abs	0.112245
arts-prons	6	fixed	raw	0.111111
arts-prons	6	fixed	rel	0.111111
arts-prons	5	fixed	abs	0.10941
arts-prons	5	fixed	raw	0.10941
arts-prons	5	fixed	rel	0.10941
arts-prons	6	open	abs	0.108844
zero	6	fixed	abs	0.108844

Simple Rank Scores based on Top 9 Usable Lines from 'date' Sample

Stopword List	No. of Links	Span Type	Link Type	Score
top100	3	fixed	raw	0.108277
top150	3	fixed	raw	0.108277
bt	3	fixed	raw	0.10771
zero	6	fixed	raw	0.107143
zero	6	fixed	rel	0.106576
top150	4	open	raw	0.105442
top100	4	open	raw	0.104875
top100	6	open	raw	0.103741
top150	6	open	raw	0.103741
top150	4	open	rel	0.103175
bt	5	open	raw	0.101474
top100	4	open	rel	0.101474
top150	5	open	raw	0.101474
top150	5	open	rel	0.101474
top50	6	open	raw	0.101474
top100	5	open	rel	0.10034
top100	5	open	raw	0.0997732
top50	5	open	rel	0.0986395
bt	4	open	rel	0.0963719
top50	3	fixed	rel	0.0946712
top50	5	open	raw	0.0946712
bt	4	open	raw	0.0918367
top50	3	open	abs	0.0901361
bt	3	fixed	abs	0
bt	3	fixed	rel	0
bt	3	open	abs	0
bt	4	fixed	abs	0
bt	4	fixed	raw	0
bt	4	fixed	rel	0
bt	4	open	abs	0
bt	5	fixed	abs	0
bt	5	fixed	raw	0
bt	5	fixed	rel	0
bt	5	open	abs	0
bt	5	open	rel	0
bt	6	fixed	abs	0
bt	6	fixed	raw	0
bt	6	fixed	rel	0
bt	6	open	abs	0
bt	6	open	raw	0
bt	6	open	rel	0
btb	4	fixed	abs	0
btb	4	fixed	raw	0
btb	4	fixed	rel	0
btb	4	open	abs	0
btb	5	fixed	abs	0
btb	5	fixed	raw	0
btb	5	fixed	rel	0
btb	5	open	abs	0
btb	5	open	rel	0
btb	6	fixed	abs	0

Appendix 5c

Simple Rank Scores based on Top 9 Usable Lines from 'date' Sample

Stopword List	No. of Links	Span Type	Link Type	Score
btb	6	fixed	raw	0
btb	6	fixed	rel	0
btb	6	open	abs	0
btb	6	open	raw	0
btb	6	open	rel	0
top100	3	fixed	abs	0
top100	3	fixed	rel	0
top100	3	open	abs	0
top100	4	fixed	abs	0
top100	4	fixed	raw	0
top100	4	fixed	rel	0
top100	4	open	abs	0
top100	5	fixed	abs	0
top100	5	fixed	raw	0
top100	5	fixed	rel	0
top100	5	open	abs	0
top100	6	fixed	abs	0
top100	6	fixed	raw	0
top100	6	fixed	rel	0
top100	6	open	abs	0
top100	6	open	rel	0
top150	3	fixed	abs	0
top150	3	fixed	rel	0
top150	3	open	abs	0
top150	4	fixed	abs	0
top150	4	fixed	raw	0
top150	4	fixed	rel	0
top150	4	open	abs	0
top150	5	fixed	abs	0
top150	5	fixed	raw	0
top150	5	fixed	rel	0
top150	5	open	abs	0
top150	6	fixed	abs	0
top150	6	fixed	raw	0
top150	6	fixed	rel	0
top150	6	open	abs	0
top150	6	open	rel	0
top50	3	fixed	abs	0
top50	4	fixed	abs	0
top50	4	fixed	raw	0
top50	4	fixed	rel	0
top50	4	open	abs	0
top50	5	fixed	abs	0
top50	5	fixed	raw	0
top50	5	fixed	rel	0
top50	5	open	abs	0
top50	6	fixed	abs	0
top50	6	fixed	raw	0
top50	6	fixed	rel	0
top50	6	open	abs	0
top50	6	open	rel	0

Appendix 6
Best-match Concordances in Evaluation

The figures attached to each line represent firstly the original line number and secondly the number of bonds.

a) Best-match Concordance Line Selection for Simple Ranking vs Representative:
arts-prons/2/open/abs

```
168 1688  be different. <LTH> As we come up to date, people # do it'' to be the same. Lo
177 1651 ning my 'firm grasp of the most up-to date trauma procedures". <t> The referenc
156 1629  to enable the returner to keep up to date with developments.<t> Various other
157 1617 ght of her man # <SO> Very much up to date, only been in service with our own f
154 1615  he said he would bring Mr Bush up to date on the issue:If we were forced to re
152 1581 tary of.state, brought Franklin up to date on the bloodshed in his beloved Fran
159 1550  so they can keep their members up to date with what is happening in the indust
188 1548 d, and though this is the first up-to date survey of its politics, it does not
153 1512  adequate nuclear weapons, kept up to date and based forward in Europe, our def
173 1495 eally because er it's just been up-to date and it <M01> Mhm.<M02> I mean that's
176 1487  now available but ask somebody up-to date.<M01> Mm.<F01> And of course compute
174 1483 ndia where we've got reasonably up to date statistics on population. Er there's
164 1461 ile, the multi-national menu is up-to date without being trendy: strikingly fre
184 1447 n though the home loan was paid up to date. Few would ever have imagined they c
158 1446 es with all the service records up to date. Abandoned only because arthritis ha
 88  615 1. Just as computers overwrite out-of date files on their disks, monks used to
 86  613 9) about British tennis are as out of date as the Dunlop Maxply in the attic. <
 83  612  the S.E.2000 would be already out of date even before it first flew, and a new
 84  604 ticisms that their magazine is out of date or has lost its edge # Editor Zanne
 89  603 ng data from instruments years out of date. Small craft allow the use of up-to-
 90  600 ed administration,traditional, out-of date, a group of elderly men smoking ciga
 85  599 F0> a newspaper it is slightly out of date but erm <F0X> Anything at all it'll
 82  598 ng them, is discriminating and out of date?They have had some support from lead
 81  595 al government were swiftly put out of date yesterday by the President of Kazakh
 87  590 formation was always six hours out of date. I get an update from the senior for
163  571 teach has not been very productive to date, nor is it likely to become more so
189  452 akes it the No. 1 film of the year to date and the biggest April release in the
187  436 s, and to considerable depths, but to date no detailed studies have been made o
 75  436 inistry denies there is a hold-up, no date has been set for a new round of talk
175  434 ties in which you've been involved to date?<M01> Er the spectrum of that would
 66  406 d by September 1993 and at some later date the US authorities will declare the
 67  401 ople will return to church at a later date.I would like to invite everyone to a
161  376  In New Cross # their greatest hit to date, is nowhere to be seen; but they do
 79  365 t letting us go to work.<t> Dugan: No date has been set for the resumption of c
 35  365  Street, London ec88 2NG. <t> Closing date for the contest is January 7, 1992.
165  353 H> Sir-We are almost there, having to date raised <KPD> 92,000 in aid of lifebo
 76  335 population is a minority in Serbia.No date has been set for elections and there
 65  325 t the same operation again at a later date. He may even, some analysts say,risk
180  319 s Milosevic's only live appearance to date came on one of its interview shows #
179  316 ents. In one of the few such moves to date, KGF recently moved the management o
132  312 ing my face again when he came to the date of birth, turning to the back to see
  6  302 d declared their willingness to set a date for starting stage two of economic a
191  291 t of ownership.<t> Levinson: No trial date has been set yet for the Janis lawsu
 57  276 iods of deep loneliness and grief.Her date of birth has been placed somewhere a
116  274  last night to set 2000 as the target date for stabilising emissions of carbon
181  263 11 potential has not been realised to date owing to the ground.<t> In the Templ
186  260 3 months; none has become infected to date. In more than 70 incidents worldwide
 93  255 he next three weeks # The NBL cut-off date for the finalisation of imports is n
```

```
 38  254 tel (about $1.2 million # The closing date for inclusion of properties is July
 30  253 rawn will each receive a kit. Closing date for entries is August 14. Standard r
 27  251 ct entries selected after the closing date of Tuesday, August 10 will win the n
 28  244 he first name drawn after the closing date on October 22, will receive a free L
169  239  financial climate in the country. To date I have only received five applicatio
166  233  notable two year-old performances to date # writes Dean Bailey <LTH> RESPONDIN
104  223 res of Winners on June 20, the record date for Friday's special meeting to cons
 32  222 as soon as possible after the closing date. <LTH> 9 Send entries to: AP/Image H
167  220 ee billion bases, or coding units. To date, fossilised DNA has been extracted f
172  219 rhythm, 'Love # is his biggest hit to date. Chang's brand of lyrics label him a
119  217  hatching of meadow birds. After that date the mechanical cultivation of fields
155  214 u work in the city?" and so on.<t> To date no machine has successfully fooled a
150  205 endence, and Chart 91 is set for this date for Helsinki, the capital, for 12.00
182  196 iminalising of breaches in the law to date.<t> The ruling Christian Democrats a
178  194  we are encouraged by our progress to date." In New York, John S. Reidy, analys
199  183 r after ordering the reactors # which date back to the 1950s # to be shut. They
 37  183 an 1500 immediately after the closing date # DAMIEN MARSH .. hard work has paid
 25  179 words and so we will see that sell-by date is no longer associated with perisha
 74  177 to finish the record, before the next date of the tour in Lisbon. <LTH> The Edg
 55  174 s had slept with a man on their first date and 39 per cent admitted to being un
 34  169  Kings Road, London sw10 OTE. Closing date is August 13. Normal rules apply. <h
 97  158  page May 3). Federal credit programs date back to the New Deal, and were meant
170  155  the most impassioned Vedder vocal to date. He creates an opening mood of lonel
 56  155 greement with the Chinese on a formal date for the resumption of diplomatic tie
121  152 e Council by 1 April 1993. After that date, it will be an offence to run an unr
162  151 heir good work. Their achievements to date are quite amazing: land reclaimed; g
 52  149  Dublin on Saturday should set a firm date for an inter-governmental conference
198  147 we started restoring our murals which date back to the Portuguese era in the 14
130  146 t Tunisia has implicitly accepted the date of the 27th, so that would suggest t
 42  145 tland Yard says the children's deaths date back to 1984. Reports suggest that b
 10  141  Almost half thought she should set a date for stepping down; 35 per cent that
143  140 is expected to confirm April 9 as the date of the election. <h> Tories pin elec
185  137 sible # says Bremner. The evidence to date, he says,suggests that men given a c
101  137  These measurements also give a quick date for that segment of the whole ice co
 31  131 nner to be notified by phone. Closing date? box 29661 <LTH> CZECH 42, passive,
183  129 as won almost all their encounters to date.<t> Short emerged from the candidate
 29  129 d Street, London se1 9LS. The closing date is Tuesday August 31, 1993 and the e
122  126 he first use of their cards from that date. They have until March 1, 1993, to c
 94  124 lained that Iraq is offering only one date for a meeting, while he has offered
 51  124 k and Barbara Sinatra after the final date of Mr Sinatra's London season. Today
 41  118  MX.<MOX> Er communications. The copy date for the next issue of Foreword is th
 80  116 others.<t> 1 Indefinite exclusion: no date is fixed for a return. Consideration
  4  116 nue to try to get you into bed. <LTH> Date rape is at the forefront of all our
123  112 ded Tuesday to have a meeting on that date, the judge ordered the meeting held
151  111  Edition # I'm Neal Conan.<t> On this date in 1956 the Republican Party nominat
124  106 ver. <CQ0> <t> Leap-horn gave him the date of the death of Pointed Shoes.<t> <C
 15  106 id Hogan, SM, has been adjourned to a date to be fixed.<dt> 930414 </dt> Cairns
142  105  Thatcher acted, bringing forward the date for a possible leadership election i
 50  105 s. <LTH> Make sure you know the final date for accepting a place. Decline unwan
 33  103 to reach us no later than the closing date, July 31, 1993. <LTH> 1. TOGECAT <LT
127  101 s <CES> <t> So it would seem from the date of his birth # <t> My God # <t> He's
117   96 oard and set an implementation target date of January 1. <t> The working party
126   92 e from the trial judge announcing the date of his execution in six weeks and on
102   89  the 'artwork' for him at the Reading date), going on to waltz until dawn with
```

```
 96   86 erly revised at the earliest possible date. <LTH> It is also unfortunate that a
171   84 nt clubs, this is Brainiak's story to date # B Sides'' features three God-bless
 13   80  and flour when he stood her up for a date. But what has been the nature of the
 12   77 t of water # At one point he forgot a date that was sort of a simple date on wh
 39   74  call 008 812 772 for details.CLOSING DATE: August 13.DRAWN: August 20.MOBILE P
125   67  was a model patient, remembering the date of every appointment and following a
 11   67 her husband like to sometimes go on a date and spend the night in a hotel.Mrs.
140   65 n the--on the--petrified tree and the date # And in all of my trips out to Mont
  7   65 o longer enjoys the preparation for a date. 'Getting ready is part of the fun o
137   64 t and your order # with a note of the date you sent it. Don't forget to give yo
106   62 0> <t> Mark Keenan, 28, whose release date had been delayed by 28 days, was fou
  9   62  they flew to Fort Worth to perform a date at the Dallas Hilton.Tina was wearin
100   59 nd FX that therefore the official pub date for the U K <FOX> <ZGY> <FOX> would
 43   58 ears ahead of its intended deployment date. <t> Only ten navigators had been tr
 20   57 however, with the addition of another date at the London Camden Falcon on Septe
145   56 dy else appears to have forgotten the date. Others feel the need to discuss the
129   55 ard, Leo asked her for a date,and the date led to this. This deal has to be cas
 59   53 ishers, July 26 is the most important date in the year. It marks the anniversar
111   51 ticle ('Crime made easy') of the same date seems to have that problem. Can you
 98   51 u do you have a sort of a prospective date for having the whole thing up and ru
 71   51 ogressive-punks play a one-off London date with Poisoned Electrick Head at New
 58   50  that it is vital.<p> Dr Salk's ideas date back a long way, but he has linked t
160   49 tributed just less than dollars 3m to date. Most big state campaigns cost about
147   49 w. Leah claimed she knew by the third date that she wanted to become Mrs Winter
146   48 ssembly. The elections, such as their date and the voters' roll and even that m
 17   46 e given here is set for this time and date, and for the capital, Paramaribo.<t>
138   45  about 13 # I can't remember what the date on that is--about 1773 or so # She--
 69   44 lf grows stronger, even if the likely date seems to recede towards the edge of
 63   44 ER: Sue Waldram PRODUCER: Ferri Jahed DATE REC: 10 July 1990 TAPE NO: 90r/32k/0
 60   41 as sensible of the priority of one in date. It was AD 450, that they beat the S
 44   41 he new Germany has chosen a different date # October 3 # Reunification Day to b
114   40  a visit with my son on such and such date else I would have been there. Probat
112   39 follow their previously announced six date tour, are priced <KPD> 8.50. Fans wi
103   38 rew it and began again, with a record date of Aug. 29. Amdura has challenged th
 54   38 > Do you play it cool after the first date? <LTH> Sarah If it was left that we
 53   38 d--don't pick Red Lobster for a first date. Great lunch deals. Hours: Mon.-Thur
196   36 ing to block a money-spinning Wembley date.Edwards hopes to convince FA Cup sem
 62   35 960, pp. 181-8).An early inauguration date for the material product concept is
 72   34 10 cent a share # There is no meeting date set as yet.Pacarc said the issue to
 14   34  to America for the Brando film and a date she wants to keep with Michelle Pfei
144   33  debate the issue.<t> But setting the date is seen as little more than a pallia
131   33 ccompanied by a printed report of the date, time and number of the attempted co
110   33 our chance of life was someone else's date with </h> death?;Steve Hyett;Part 2
 61   33 [heb.] shows that the poem is late in date. However,Phoenician inscriptions ear
 48   33 of parties in parties in the Election date election parliament Albania Mar 23rd
 19   33 . <LTH> JULIAN COPE has added another date to his ''Head On'' tour. It's at Bra
 77   32 ention a place. And so far there's no date fixed for the meeting. Until that ha
  1   32 LTH> Opening up for Metallica on a 65 date US tour, The Cult banged on with the
141   31  intervention until September 20, the date of the referendum, if necessary, and
 95   31 xhaustion of a new mother. An opening date in June would have given her two pre
 49   31 -partisan. In announcing the election date, President Roh Tae Woo said there wa
139   30 d war against Iraq # He also said the date is imminent and that a ground war ca
 78   30 e, no proof, no dossier, no names, no date, no body.And as happens in all hosta
120   28 wever, always say that he shares that date with my wife. <LTH> Dr L Keith Franc
```

194 27 occur in the group of kouroi that we date the earliest.<t> JAFFE: The torso, o
190 27 here tier upon tier of rock-cut tombs date from as long ago as the thirteenth c
16 27 as been lined up as his dating agency date unbeknown to wife Alison Steadman wh
36 25 ulled from our postbag on the closing date. <t> Over-18's only. News Internatio
200 23 lossomed in the presence of women who date act ors and princes, dine in Milan a
105 23 0> Twenty years on from their release date, two albums look set to make this mo
118 22 ns has two weeks from his termination date to appeal the decision. As for Randa
99 22 he Autumn Triangle. <LTH> Provisional date and venue for National Council 1992:
193 21 the announced 20 March maiden voyage date. On 10 October the company released
128 21 ork on a regular basis and is not the date his family or household goods and ef
45 20 y 8. You may get a slightly different date by this short cut method than by add
133 18 like the driver of a Hansom cab. The date is 29 March 1920.<t> <FCH> Above lef
91 18 a puritan streak, and the concept of date" or 'acquaintance" rape reveals just
26 18 arded for decades, only an expert can date a garment. When a skirt length chang
2 18 d </pres> <prod> Francis Mead </prod> Date Rec: 16 October 1990 Prog No: 90r/32
148 17 e asks you out again despite no third date action), you know you've built a fou
197 16 It's not stated clearly back to what date this is effective # The decrees come
135 16 n the exact calendar months after the date the loan was opened. <LTH> Written q
108 15 the 11th hour cancellation of Suede's date at the venue, and the closure of the
113 14 osed October the 30th as the starting date # Mrs # Mandela's lawyer argued succ
46 13 egin to accumulate with each dividend date. drps really do serve an important f
22 13 eels, and as an adult his first blind date.<t> Unidentified Woman (From Radio A
136 12 ou post your order and payment by the date on the enclosed form. <LTH> <FCH> Bu
40 12 opening rounds. A mutually convenient date should then be set and green fees sh
68 9 nd its disposition, and of the likely date when the accumulated treasure, with
192 8 ane 2222 No built <FCH> Aircraft Type Date Purpose of Design No built <FCH> Ans
115 8 le, the village is surrounded by tall date palms and lush green farmland. Its n
47 8 tion figure said the distant election date will give the ruling family time to
23 8 ed banner headlines about his # blind date'' escapade. <t> It was that sense of
18 8 he next regularly slated announcement date is Sept. 18. A brand new directorate
73 7 s portrayed.'' <t> <h> Benetton's new date;Motor Racing </h> <dt> 25 August 199
24 6 LISTINGS <LTH> Concerts are listed by date, then by city # Classical Listings c
149 5 London,1981). Berlitz associated this date with the dire predictions given in G
70 5 ks (S) on design problems (location # date).Contract for Snabl already done, ps
8 5 ance Group, who will present We Got A Date, Can't Take Johnny To The Funeral an
3 5 details of each match you play. <LTH> DATE COMPETITION OPPONENT VENUE SCORE RES
195 4 inals in Filderstadt. <t> <h> Wembley date;Rugby League </h> <dt> 15 October 19
109 4 ally signed and predated with today's date), his eyeglasses, a Koran, a Bible.
134 3 rmehrung beim Umbau," which bears the date December 13, 1932 at the end.39 Ibid
107 3 n vaults and galleries. Those in Rome date mainly from the third and early four
21 3 tion ready for critics of the bizarre date-rape story, 'What Actually Happened
64 2 r, Sir Andrew Lloyd Webber. <t> Janet Date, a guide and former actress, is in h
5 1 Schedule </h> Playing this weekend:A Date With Judy (1948) Jane Powell plays a
92 0 ers announced the April 5th blast-off date following a flight review at the Ken

b) Best-match Concordance Line Selection for Simple Ranking vs Usable: arts-prons/2/fixed/abs

```
157 1512 ght of her man # <SO> Very much up to date, only been in service with our own f
159 1494  so they can keep their members up to date with what is happening in the indust
156 1471  to enable the returner to keep up to date with developments.<t> Various other
152 1464 tary of state, brought Franklin up to date on the bloodshed in his beloved Fran
168 1454  be different. <LTH> As we come up to date, people # do it'' to be the same. Lo
176 1451  now available but ask somebody up-to date.<M01> Mm.<F01> And of course compute
173 1449 eally because er it's just been up-to date and it <M01> Mhm.<M02> I mean that's
188 1438 d, and though this is the first up-to date survey of its politics, it does not
184 1436 n though the home loan was paid up to date. Few would ever have imagined they c
154 1435  he said he would bring Mr Bush up to date on the issue:If we were forced to re
164 1433 ile, the multi-national menu is up-to date without being trendy: strikingly fre
177 1432 ning my 'firm grasp of the most up-to date trauma procedures". <t> The referenc
153 1431  adequate nuclear weapons, kept up to date and based forward in Europe, our def
174 1414 ndia where we've got reasonably up to date statistics on population. Er there's
158 1414 es with all the service records up to date. Abandoned only because arthritis ha
 86  600 9) about British tennis are as out of date as the Dunlop Maxply in the attic. <
 84  594 ticisms that their magazine is out of date or has lost its edge # Editor Zanne
 85  593 F0> a newspaper it is slightly out of date but erm <F0X> Anything at all it'll
 88  591 1. Just as computers overwrite out-of date files on their disks, monks used to
 83  591  the S.E.2000 would be already out of date even before it first flew, and a new
 82  590 ng them, is discriminating and out of date?They have had some support from lead
 90  589 ed administration,traditional, out-of date, a group of elderly men smoking ciga
 81  588 al government were swiftly put out of date yesterday by the President of Kazakh
 89  587 ng data from instruments years out of date. Small craft allow the use of up-to-
 87  587 formation was always six hours out of date. I get an update from the senior for
 66  345 d by September 1993 and at some later date the US authorities will declare the
 79  317 t letting us go to work.<t> Dugan: No date has been set for the resumption of c
 75  313 inistry denies there is a hold-up, no date has been set for a new round of talk
 67  310 ople will return to church at a later date.I would like to invite everyone to a
 76  309 population is a minority in Serbia.No date has been set for elections and there
 35  300  Street, London ec88 2NG. <t> Closing date for the contest is January 7, 1992.
 65  296 t the same operation again at a later date. He may even, some analysts say,risk
191  257 t of ownership.<t> Levinson: No trial date has been set yet for the Janis lawsu
132  250 ing my face again when he came to the date of birth, turning to the back to see
  6  233 d declared their willingness to set a date for starting stage two of economic a
189  224 akes it the No. 1 film of the year to date and the biggest April release in the
 30  219 rawn will each receive a kit. Closing date for entries is August 14. Standard r
161  217  In New Cross # their greatest hit to date, is nowhere to be seen; but they do
 38  215 tel (about $1.2 million # The closing date for inclusion of properties is July
 57  214 iods of deep loneliness and grief.Her date of birth has been placed somewhere a
 27  213 ct entries selected after the closing date of Tuesday, August 10 will win the n
 28  202 he first name drawn after the closing date on October 22, will receive a free L
116  189  last night to set 2000 as the target date for stabilising emissions of carbon
180  184 s Milosevic's only live appearance to date came on one of its interview shows #
 32  180 as soon as possible after the closing date. <LTH> 9 Send entries to: AP/Image H
 34  166  Kings Road, London sw10 OTE. Closing date is August 13. Normal rules apply. <h
 37  164 an 1500 immediately after the closing date # DAMIEN MARSH .. hard work has paid
175  163 ties in which you've been involved to date?<M01> Er the spectrum of that would
 25  145 words and so we will see that sell-by date is no longer associated with perisha
198  143 we started restoring our murals which date back to the Portuguese era in the 14
 97  142  page May 3). Federal credit programs date back to the New Deal, and were meant
```

42 142 tland Yard says the children's deaths date back to 1984. Reports suggest that b
199 141 r after ordering the reactors # which date back to the 1950s # to be shut. They
186 138 3 months; none has become infected to date. In more than 70 incidents worldwide
169 137 financial climate in the country. To date I have only received five applicatio
178 134 we are encouraged by our progress to date." In New York, John S. Reidy, analys
93 134 he next three weeks # The NBL cut-off date for the finalisation of imports is n
55 132 s had slept with a man on their first date and 39 per cent admitted to being un
101 130 These measurements also give a quick date for that segment of the whole ice co
56 130 greement with the Chinese on a formal date for the resumption of diplomatic tie
10 121 Almost half thought she should set a date for stepping down; 35 per cent that
172 120 rhythm, 'Love # is his biggest hit to date. Chang's brand of lyrics label him a
121 117 e Council by 1 April 1993. After that date, it will be an offence to run an unr
181 116 ll potential has not been realised to date owing to the ground.<t> In the Templ
150 116 endence, and Chart 91 is set for this date for Helsinki, the capital, for 12.00
29 109 d Street, London sel 9LS. The closing date is Tuesday August 31, 1993 and the e
165 102 H> Sir-We are almost there, having to date raised <KPD> 92,000 in aid of lifebo
167 100 ee billion bases, or coding units. To date, fossilised DNA has been extracted f
163 99 teach has not been very productive to date, nor is it likely to become more so
155 92 u work in the city?" and so on.<t> To date no machine has successfully fooled a
119 88 hatching of meadow birds. After that date the mechanical cultivation of fields
51 86 k and Barbara Sinatra after the final date of Mr Sinatra's London season. Today
187 85 s, and to considerable depths, but to date no detailed studies have been made o
52 85 Dublin on Saturday should set a firm date for an inter-governmental conference
33 83 to reach us no later than the closing date, July 31, 1993. <LTH> 1. TOGECAT <LT
151 81 Edition # I'm Neal Conan.<t> On this date in 1956 the Republican Party nominat
15 81 id Hogan, SM, has been adjourned to a date to be fixed.<dt> 930414 </dt> Cairns
124 80 ver. <CQ0> <t> Leap-horn gave him the date of the death of Pointed Shoes.<t> <C
94 79 lained that Iraq is offering only one date for a meeting, while he has offered
127 77 s <CES> <t> So it would seem from the date of his birth # <t> My God # <t> He's
123 69 ded Tuesday to have a meeting on that date, the judge ordered the meeting held
182 68 iminalising of breaches in the law to date.<t> The ruling Christian Democrats a
162 64 heir good work. Their achievements to date are quite amazing: land reclaimed; g
117 61 oard and set an implementation target date of January 1. <t> The working party
183 59 as won almost all their encounters to date.<t> Short emerged from the candidate
39 54 call 008 812 772 for details.CLOSING DATE: August 13.DRAWN: August 20.MOBILE P
185 52 sible # says Bremner. The evidence to date, he says,suggests that men given a c
171 51 nt clubs, this is Brainiak's story to date # B Sides'' features three God-bless
106 50 0> <t> Mark Keenan, 28, whose release date had been delayed by 28 days, was fou
80 44 others.<t> 1 Indefinite exclusion: no date is fixed for a return. Consideration
74 44 to finish the record, before the next date of the tour in Lisbon. <LTH> The Edg
13 40 and flour when he stood her up for a date. But what has been the nature of the
111 39 ticle ('Crime made easy') of the same date seems to have that problem. Can you
143 36 is expected to confirm April 9 as the date of the election. <h> Tories pin elec
53 36 d--don't pick Red Lobster for a first date. Great lunch deals. Hours: Mon.-Thur
122 34 he first use of their cards from that date. They have until March 1, 1993, to c
17 31 e given here is set for this time and date, and for the capital, Paramaribo.<t>
71 30 ogressive-punks play a one-off London date with Poisoned Electrick Head at New
104 27 res of Winners on June 20, the record date for Friday's special meeting to cons
103 26 rew it and began again, with a record date of Aug. 29. Amdura has challenged th
77 26 ention a place. And so far there's no date fixed for the meeting. Until that ha
63 26 ER: Sue Waldram PRODUCER: Ferri Jahed DATE REC: 10 July 1990 TAPE NO: 90r/32k/0
41 26 MX.<M0X> Er communications. The copy date for the next issue of Foreword is th
98 25 u do you have a sort of a prospective date for having the whole thing up and ru
141 24 intervention until September 20, the date of the referendum, if necessary, and

126	23	e from the trial judge announcing the	date of his execution in six weeks and on
96	23	erly revised at the earliest possible	date. <LTH> It is also unfortunate that a
59	23	ishers, July 26 is the most important	date in the year. It marks the anniversar
20	22	however, with the addition of another	date at the London Camden Falcon on Septe
19	22	. <LTH> JULIAN COPE has added another	date to his ''Head On'* tour. It's at Bra
31	20	nner to be notified by phone. Closing	date? box 29661 <LTH> CZECH 42, passive,
190	19	here tier upon tier of rock-cut tombs	date from as long ago as the thirteenth c
130	19	t Tunisia has implicitly accepted the	date of the 27th, so that would suggest t
120	19	wever, always say that he shares that	date with my wife. <LTH> Dr L Keith Franc
54	19	> Do you play it cool after the first	date? <LTH> Sarah If it was left that we
147	18	w. Leah claimed she knew by the third	date that she wanted to become Mrs Winter
131	18	ccompanied by a printed report of the	date, time and number of the attempted co
72	18	10 cent a share # There is no meeting	date set as yet.Pacarc said the issue to
12	18	t of water # At one point he forgot a	date that was sort of a simple date on wh
160	17	tributed just less than dollars 3m to	date. Most big state campaigns cost about
138	17	about 13 # I can't remember what the	date on that is--about 1773 or so # She--
45	16	y 8. You may get a slightly different	date by this short cut method than by add
4	15	nue to try to get you into bed. <LTH>	Date rape is at the forefront of all our
95	14	xhaustion of a new mother. An opening	date in June would have given her two pre
142	13	Thatcher acted, bringing forward the	date for a possible leadership election i
91	13	a puritan streak, and the concept of	date" or 'acquaintance" rape reveals just
2	13	d </pres> <prod> Francis Mead </prod>	Date Rec: 16 October 1990 Prog No: 90r/32
166	12	notable two year-old performances to	date # writes Dean Bailey <LTH> RESPONDIN
102	12	the 'artwork' for him at the Reading	date), going on to waltz until dawn with
22	12	eels, and as an adult his first blind	date.<t> Unidentified Woman (From Radio A
144	11	debate the issue.<t> But setting the	date is seen as little more than a pallia
36	11	ulled from our postbag on the closing	date. <t> Over-18's only. News Internatio
50	10	s. <LTH> Make sure you know the final	date for accepting a place. Decline unwan
40	10	opening rounds. A mutually convenient	date should then be set and green fees sh
135	9	n the exact calendar months after the	date the loan was opened. <LTH> Written q
114	9	a visit with my son on such and such	date else I would have been there. Probat
110	9	our chance of life was someone else's	date with </h> death?;Steve Hyett;Part 2
43	9	ears ahead of its intended deployment	date. <t> Only ten navigators had been tr
9	9	they flew to Fort Worth to perform a	date at the Dallas Hilton.Tina was wearin
146	8	ssembly. The elections, such as their	date and the voters' roll and even that m
136	8	ou post your order and payment by the	date on the enclosed form. <LTH> <FCH> Bu
113	8	osed October the 30th as the starting	date # Mrs # Mandela's lawyer argued succ
1	8	LTH> Opening up for Metallica on a 65	date US tour, The Cult banged on with the
99	7	he Autumn Triangle. <LTH> Provisional	date and venue for National Council 1992:
11	7	her husband like to sometimes go on a	date and spend the night in a hotel.Mrs.
193	6	the announced 20 March maiden voyage	date. On 10 October the company released
139	6	d war against Iraq # He also said the	date is imminent and that a ground war ca
100	6	nd FX that therefore the official pub	date for the U K <FOX> <ZGY> <FOX> would
49	6	-partisan. In announcing the election	date, President Roh Tae Woo said there wa
197	5	It's not stated clearly back to what	date this is effective # The decrees come
129	5	ard, Leo asked her for a date,and the	date led to this. This deal has to be cas
69	5	lf grows stronger, even if the likely	date seems to recede towards the edge of
24	5	LISTINGS <LTH> Concerts are listed by	date, then by city # Classical Listings c
145	4	dy else appears to have forgotten the	date. Others feel the need to discuss the
128	4	ork on a regular basis and is not the	date his family or household goods and ef
115	4	le, the village is surrounded by tall	date palms and lush green farmland. Its n
112	4	follow their previously announced six	date tour, are priced <KPD> 8.50. Fans wi
108	4	the 11th hour cancellation of Suede's	date at the venue, and the closure of the
179	3	ents. In one of the few such moves to	date, KGF recently moved the management o

```
137    3 t and your order # with a note of the date you sent it. Don't forget to give yo
133    3  like the driver of a Hansom cab. The date is 29 March 1920.<t> <FCH> Above lef
125    3  was a model patient, remembering the date of every appointment and following a
118    3 ns has two weeks from his termination date to appeal the decision. As for Randa
105    3 0> Twenty years on from their release date, two albums look set to make this mo
 21    3 tion ready for critics of the bizarre date-rape story, 'What Actually Happened
 14    3  to America for the Brando film and a date she wants to keep with Michelle Pfei
  7    3 o longer enjoys the preparation for a date. 'Getting ready is part of the fun o
200    2 lossomed in the presence of women who date act ors and princes, dine in Milan a
195    2 inals in Filderstadt. <t> <h> Wembley date;Rugby League </h> <dt> 15 October 19
194    2  occur in the group of kouroi that we date the earliest.<t> JAFFE: The torso, o
192    2 ane 2222 No built <FCH> Aircraft Type Date Purpose of Design No built <FCH> Ans
140    2 n the--on the--petrified tree and the date # And in all of my trips out to Mont
107    2 n vaults and galleries. Those in Rome date mainly from the third and early four
 78    2 e, no proof, no dossier, no names, no date, no body.And as happens in all hosta
 61    2 [heb.] shows that the poem is late in date. However,Phoenician inscriptions ear
 60    2 as sensible of the priority of one in date. It was AD 450, that they beat the S
 18    2 he next regularly slated announcement date is Sept. 18. A brand new directorate
196    1 ing to block a money-spinning Wembley date.Edwards hopes to convince FA Cup sem
170    1  the most impassioned Vedder vocal to date. He creates an opening mood of lonel
 70    1 ks (S) on design problems (location # date).Contract for Snabl already done, ps
 62    1 960, pp. 181-8).An early inauguration date for the material product concept is
 58    1  that it is vital.<p> Dr Salk's ideas date back a long way, but he has linked t
 48    1 of parties in parties in the Election date election parliament Albania Mar 23rd
 44    1 he new Germany has chosen a different date # October 3 # Reunification Day to b
 23    1 ed banner headlines about his # blind date'' escapade. <t> It was that sense of
  8    1 ance Group, who will present We Got A Date, Can't Take Johnny To The Funeral an
  5    1  Schedule </h> Playing this weekend:A Date With Judy (1948) Jane Powell plays a
149    0 London,1981). Berlitz associated this date with the dire predictions given in G
148    0 e asks you out again despite no third date action), you know you've built a fou
134    0 rmehrung beim Umbau," which bears the date December 13, 1932 at the end.39 Ibid
109    0 ally signed and predated with today's date), his eyeglasses, a Koran, a Bible.
 92    0 ers announced the April 5th blast-off date following a flight review at the Ken
 73    0 s portrayed.'' <t> <h> Benetton's new date;Motor Racing </h> <dt> 25 August 199
 68    0 nd its disposition, and of the likely date when the accumulated treasure, with
 64    0 r, Sir Andrew Lloyd Webber. <t> Janet Date, a guide and former actress, is in h
 47    0 tion figure said the distant election date will give the ruling family time to
 46    0 egin to accumulate with each dividend date. drps really do serve an important f
 26    0 arded for decades, only an expert can date a garment. When a skirt length chang
 16    0 as been lined up as his dating agency date unbeknown to wife Alison Steadman wh
  3    0 details of each match you play. <LTH> DATE COMPETITION OPPONENT VENUE SCORE RES
```

c) Best-match Concordance Line Selection for Pearson Correlation 1 vs Representative:

arts-prons/4/open/rel

```
 75  484 inistry denies there is a hold-up, no date has been set for a new round of talk
163  479 teach has not been very productive to date, nor is it likely to become more so
 79  472 t letting us go to work.<t> Dugan: No date has been set for the resumption of c
161  444  In New Cross # their greatest hit to date, is nowhere to be seen; but they do
157  424 ght of her man # <SO> Very much up to date, only been in service with our own f
168  410  be different. <LTH> As we come up to date, people # do it'' to be the same. Lo
 76  408 population is a minority in Serbia.No date has been set for elections and there
175  387 ties in which you've been involved to date?<M01> Er the spectrum of that would
188  301 d, and though this is the first up-to date survey of its politics, it does not
191  282 t of ownership.<t> Levinson: No trial date has been set yet for the Janis lawsu
119  281  hatching of meadow birds. After that date the mechanical cultivation of fields
159  270  so they can keep their members up to date with what is happening in the indust
189  269 akes it the No. 1 film of the year to date and the biggest April release in the
152  234 tary of state, brought Franklin up to date on the bloodshed in his beloved Fran
 74  231 to finish the record, before the next date of the tour in Lisbon. <LTH> The Edg
 93  212 he next three weeks # The NBL cut-off date for the finalisation of imports is n
177  205 ning my 'firm grasp of the most up-to date trauma procedures". <t> The referenc
165  198 H> Sir-We are almost there, having to date raised <KPD> 92,000 in aid of lifebo
154  179  he said he would bring Mr Bush up to date on the issue:If we were forced to re
 30  163 rawn will each receive a kit. Closing date for entries is August 14. Standard r
 12  152 t of water # At one point he forgot a date that was sort of a simple date on wh
  6  148 d declared their willingness to set a date for starting stage two of economic a
 27  146 ct entries selected after the closing date of Tuesday, August 10 will win the n
153  138  adequate nuclear weapons, kept up to date and based forward in Europe, our def
156  136  to enable the returner to keep up to date with developments.<t> Various other
179  134 ents. In one of the few such moves to date, KGF recently moved the management o
 28  133 he first name drawn after the closing date on October 22, will receive a free L
 15  127 id Hogan, SM, has been adjourned to a date to be fixed.<dt> 930414 </dt> Cairns
150  126 endence, and Chart 91 is set for this date for Helsinki, the capital, for 12.00
116  126  last night to set 2000 as the target date for stabilising emissions of carbon
 83  126  the S.E.2000 would be already out of date even before it first flew, and a new
140  120 n the--on the--petrified tree and the date # And in all of my trips out to Mont
102  118  the 'artwork' for him at the Reading date), going on to waltz until dawn with
 67  118 ople will return to church at a later date.I would like to invite everyone to a
 35  115  Street, London ec88 2NG. <t> Closing date for the contest is January 7, 1992.
 89  113 ng data from instruments years out of date. Small craft allow the use of up-to-
104  111 res of Winners on June 20, the record date for Friday's special meeting to cons
181  109 ll potential has not been realised to date owing to the ground.<t> In the Templ
 29  109 d Street, London sel 9LS. The closing date is Tuesday August 31, 1993 and the e
 13  103  and flour when he stood her up for a date. But what has been the nature of the
155  102 u work in the city?" and so on.<t> To date no machine has successfully fooled a
 58  102  that it is vital.<p> Dr Salk's ideas date back a long way, but he has linked t
176  101  now available but ask somebody up-to date.<M01> Mm.<F01> And of course compute
 84  101 ticisms that their magazine is out of date or has lost its edge # Editor Zanne
 86   99 9) about British tennis are as out of date as the Dunlop Maxply in the attic. <
122   98 he first use of their cards from that date. They have until March 1, 1993, to c
121   98 e Council by 1 April 1993. After that date, it will be an offence to run an unr
 80   97 others.<t> 1 Indefinite exclusion: no date is fixed for a return. Consideration
 90   95 ed administration,traditional, out-of date, a group of elderly men smoking ciga
 25   89 words and so we will see that sell-by date is no longer associated with perisha
```

17 84 e given here is set for this time and date, and for the capital, Paramaribo.<t>
146 83 ssembly. The elections, such as their date and the voters' roll and even that m
145 83 dy else appears to have forgotten the date. Others feel the need to discuss the
114 83 a visit with my son on such and such date else I would have been there. Probat
174 82 ndia where we've got reasonably up to date statistics on population. Er there's
143 82 is expected to confirm April 9 as the date of the election. <h> Tories pin elec
126 81 e from the trial judge announcing the date of his execution in six weeks and on
66 81 d by September 1993 and at some later date the US authorities will declare the
167 80 ee billion bases, or coding units. To date, fossilised DNA has been extracted f
57 76 iods of deep loneliness and grief.Her date of birth has been placed somewhere a
187 70 s, and to considerable depths, but to date no detailed studies have been made o
138 70 about 13 # I can't remember what the date on that is--about 1773 or so # She--
55 70 s had slept with a man on their first date and 39 per cent admitted to being un
112 69 follow their previously announced six date tour, are priced <KPD> 8.50. Fans wi
71 64 ogressive-punks play a one-off London date with Poisoned Electrick Head at New
20 64 however, with the addition of another date at the London Camden Falcon on Septe
180 59 s Milosevic's only live appearance to date came on one of its interview shows #
166 58 notable two year-old performances to date # writes Dean Bailey <LTH> RESPONDIN
151 58 Edition # I'm Neal Conan.<t> On this date in 1956 the Republican Party nominat
61 57 [heb.] shows that the poem is late in date. However,Phoenician inscriptions ear
4 55 nue to try to get you into bed. <LTH> Date rape is at the forefront of all our
132 54 ing my face again when he came to the date of birth, turning to the back to see
72 54 10 cent a share # There is no meeting date set as yet.Pacarc said the issue to
96 52 erly revised at the earliest possible date. <LTH> It is also unfortunate that a
65 52 t the same operation again at a later date. He may even, some analysts say,risk
82 51 ng them, is discriminating and out of date?They have had some support from lead
147 50 w. Leah claimed she knew by the third date that she wanted to become Mrs Winter
38 49 tel (about $1.2 million # The closing date for inclusion of properties is July
178 46 we are encouraged by our progress to date." In New York, John S. Reidy, analys
94 46 lained that Iraq is offering only one date for a meeting, while he has offered
63 46 ER: Sue Waldram PRODUCER: Ferri Jahed DATE REC: 10 July 1990 TAPE NO: 90r/32k/0
32 45 as soon as possible after the closing date. <LTH> 9 Send entries to: AP/Image H
130 44 t Tunisia has implicitly accepted the date of the 27th, so that would suggest t
33 44 to reach us no later than the closing date, July 31, 1993. <LTH> 1. TOGECAT <LT
85 43 F0> a newspaper it is slightly out of date but erm <F0X> Anything at all it'll
169 42 financial climate in the country. To date I have only received five applicatio
43 42 ears ahead of its intended deployment date. <t> Only ten navigators had been tr
164 41 ile, the multi-national menu is up-to date without being trendy: strikingly fre
60 40 as sensible of the priority of one in date. It was AD 450, that they beat the S
19 40 . <LTH> JULIAN COPE has added another date to his ''Head On'" tour. It's at Bra
186 36 3 months; none has become infected to date. In more than 70 incidents worldwide
1 36 LTH> Opening up for Metallica on a 65 date US tour, The Cult banged on with the
173 35 eally because er it's just been up-to date and it <M01> Mhm.<M02> I mean that's
111 35 ticle ('Crime made easy') of the same date seems to have that problem. Can you
105 35 0> Twenty years on from their release date, two albums look set to make this mo
81 35 al government were swiftly put out of date yesterday by the President of Kazakh
39 34 call 008 812 772 for details.CLOSING DATE: August 13.DRAWN: August 20.MOBILE P
137 33 t and your order # with a note of the date you sent it. Don't forget to give yo
14 33 to America for the Brando film and a date she wants to keep with Michelle Pfei
123 32 ded Tuesday to have a meeting on that date, the judge ordered the meeting held
184 31 n though the home loan was paid up to date. Few would ever have imagined they c
87 31 formation was always six hours out of date. I get an update from the senior for
88 30 1. Just as computers overwrite out-of date files on their disks, monks used to
59 30 ishers, July 26 is the most important date in the year. It marks the anniversar

```
  7  30 o longer enjoys the preparation for a date. 'Getting ready is part of the fun o
158  29 es with all the service records up to date. Abandoned only because arthritis ha
 77  28 ention a place. And so far there's no date fixed for the meeting. Until that ha
197  23  It's not stated clearly back to what date this is effective # The decrees come
 34  23  Kings Road, London sw10 OTE. Closing date is August 13. Normal rules apply. <h
172  22 rhythm, 'Love # is his biggest hit to date. Chang's brand of lyrics label him a
139  22 d war against Iraq # He also said the date is imminent and that a ground war ca
 54  22 > Do you play it cool after the first date? <LTH> Sarah If it was left that we
 41  22  MX.<MOX> Er communications. The copy date for the next issue of Foreword is th
 56  21 greement with the Chinese on a formal date for the resumption of diplomatic tie
 52  20  Dublin on Saturday should set a firm date for an inter-governmental conference
200  19 lossomed in the presence of women who date act ors and princes, dine in Milan a
 53  19 d--don't pick Red Lobster for a first date. Great lunch deals. Hours: Mon.-Thur
129  18 ard, Leo asked her for a date,and the date led to this. This deal has to be cas
 49  17 -partisan. In announcing the election date, President Roh Tae Woo said there wa
144  15  debate the issue.<t> But setting the date is seen as little more than a pallia
194  14  occur in the group of kouroi that we date the earliest.<t> JAFFE: The torso, o
  2  14 d </pres> <prod> Francis Mead </prod> Date Rec: 16 October 1990 Prog No: 90r/32
 31  13 nner to be notified by phone. Closing date? box 29661 <LTH> CZECH 42, passive,
142  12  Thatcher acted, bringing forward the date for a possible leadership election i
100  12 nd FX that therefore the official pub date for the U K <FOX> <ZGY> <FOX> would
 69  12 lf grows stronger, even if the likely date seems to recede towards the edge of
 11  12 her husband like to sometimes go on a date and spend the night in a hotel.Mrs.
128  11 ork on a regular basis and is not the date his family or household goods and ef
 16  11 as been lined up as his dating agency date unbeknown to wife Alison Steadman wh
190  10 here tier upon tier of rock-cut tombs date from as long ago as the thirteenth c
 98  10 u do you have a sort of a prospective date for having the whole thing up and ru
182   9 iminalising of breaches in the law to date.<t> The ruling Christian Democrats a
162   9 heir good work. Their achievements to date are quite amazing: land reclaimed; g
118   9 ns has two weeks from his termination date to appeal the decision. As for Randa
127   8 s <CES> <t> So it would seem from the date of his birth # <t> My God # <t> He's
 78   8 e, no proof, no dossier, no names, no date, no body.And as happens in all hosta
131   7 ccompanied by a printed report of the date, time and number of the attempted co
110   7 our chance of life was someone else's date with </h> death?;Steve Hyett;Part 2
 95   7 xhaustion of a new mother. An opening date in June would have given her two pre
 45   7 y 8. You may get a slightly different date by this short cut method than by add
 37   7 an 1500 immediately after the closing date # DAMIEN MARSH .. hard work has paid
183   6 as won almost all their encounters to date.<t> Short emerged from the candidate
135   6 n the exact calendar months after the date the loan was opened. <LTH> Written q
106   6 0> <t> Mark Keenan, 28, whose release date had been delayed by 28 days, was fou
 50   6 s. <LTH> Make sure you know the final date for accepting a place. Decline unwan
 36   6 ulled from our postbag on the closing date. <t> Over-18's only. News Internatio
160   5 tributed just less than dollars 3m to date. Most big state campaigns cost about
117   5 oard and set an implementation target date of January 1. <t> The working party
 99   5 he Autumn Triangle. <LTH> Provisional date and venue for National Council 1992:
185   4 sible # says Bremner. The evidence to date, he says,suggests that men given a c
 97   4  page May 3). Federal credit programs date back to the New Deal, and were meant
171   3 nt clubs, this is Brainiak's story to date # B Sides'' features three God-bless
148   3 e asks you out again despite no third date action), you know you've built a fou
141   3  intervention until September 20, the date of the referendum, if necessary, and
 70   3 ks (S) on design problems (location # date).Contract for Snabl already done, ps
 40   3 opening rounds. A mutually convenient date should then be set and green fees sh
 26   3 arded for decades, only an expert can date a garment. When a skirt length chang
 24   3 LISTINGS <LTH> Concerts are listed by date, then by city # Classical Listings c
```

```
23    3 ed banner headlines about his # blind date'' escapade. <t> It was that sense of
10    3  Almost half thought she should set a date for stepping down; 35 per cent that
 9    3  they flew to Fort Worth to perform a date at the Dallas Hilton.Tina was wearin
199   2 r after ordering the reactors # which date back to the 1950s # to be shut. They
196   2 ing to block a money-spinning Wembley date.Edwards hopes to convince FA Cup sem
193   2  the announced 20 March maiden voyage date. On 10 October the company released
192   2 ane 2222 No built <FCH> Aircraft Type Date Purpose of Design No built <FCH> Ans
124   2 ver. <CQ0> <t> Leap-horn gave him the date of the death of Pointed Shoes.<t> <C
101   2  These measurements also give a quick date for that segment of the whole ice co
 68   2 nd its disposition, and of the likely date when the accumulated treasure, with
 51   2 k and Barbara Sinatra after the final date of Mr Sinatra's London season. Today
 44   2 he new Germany has chosen a different date # October 3 # Reunification Day to b
136   1 ou post your order and payment by the date on the enclosed form. <LTH> <FCH> Bu
133   1  like the driver of a Hansom cab. The date is 29 March 1920.<t> <FCH> Above lef
125   1  was a model patient, remembering the date of every appointment and following a
120   1 wever, always say that he shares that date with my wife. <LTH> Dr L Keith Franc
115   1 le, the village is surrounded by tall date palms and lush green farmland. Its n
103   1 rew it and began again, with a record date of Aug. 29. Amdura has challenged th
 92   1 ers announced the April 5th blast-off date following a flight review at the Ken
 48   1 of parties in parties in the Election date election parliament Albania Mar 23rd
 47   1 tion figure said the distant election date will give the ruling family time to
 42   1 tland Yard says the children's deaths date back to 1984. Reports suggest that b
  3   1 details of each match you play. <LTH> DATE COMPETITION OPPONENT VENUE SCORE RES
198   0 we started restoring our murals which date back to the Portuguese era in the 14
195   0 inals in Filderstadt. <t> <h> Wembley date;Rugby League </h> <dt> 15 October 19
170   0  the most impassioned Vedder vocal to date. He creates an opening mood of lonel
149   0 London,1981). Berlitz associated this date with the dire predictions given in G
134   0 rmehrung beim Umbau," which bears the date December 13, 1932 at the end.39 Ibid
113   0 osed October the 30th as the starting date # Mrs # Mandela's lawyer argued succ
109   0 ally signed and predated with today's date), his eyeglasses, a Koran, a Bible.
108   0 the 11th hour cancellation of Suede's date at the venue, and the closure of the
107   0 n vaults and galleries. Those in Rome date mainly from the third and early four
 91   0  a puritan streak, and the concept of date" or 'acquaintance" rape reveals just
 73   0 s portrayed.'' <t> <h> Benetton's new date;Motor Racing </h> <dt> 25 August 199
 64   0 r, Sir Andrew Lloyd Webber. <t> Janet Date, a guide and former actress, is in h
 62   0 960, pp. 181-8).An early inauguration date for the material product concept is
 46   0 egin to accumulate with each dividend date. drps really do serve an important f
 22   0 eels, and as an adult his first blind date.<t> Unidentified Woman (From Radio A
 21   0 tion ready for critics of the bizarre date-rape story, 'What Actually Happened
 18   0 he next regularly slated announcement date is Sept. 18. A brand new directorate
  8   0 ance Group, who will present We Got A Date, Can't Take Johnny To The Funeral an
  5   0  Schedule </h> Playing this weekend:A Date With Judy (1948) Jane Powell plays a
```

d) Best-match Concordance Line Selection for Pearson Correlation 1 vs Usable: zero/1/fixed/abs

189 6429 akes it the No. 1 film of the year to date and the biggest April release in the
152 5590 tary of state, brought Franklin up to date on the bloodshed in his beloved Fran
154 5494 he said he would bring Mr Bush up to date on the issue:If we were forced to re
175 5416 ties in which you've been involved to date?<M01> Er the spectrum of that would
177 4835 ning my 'firm grasp of the most up-to date trauma procedures". <t> The referenc
182 4743 iminalising of breaches in the law to date.<t> The ruling Christian Democrats a
181 4577 ll potential has not been realised to date owing to the ground.<t> In the Templ
173 4482 eally because er it's just been up-to date and it <M01> Mhm.<M02> I mean that's
 74 4450 to finish the record, before the next date of the tour in Lisbon. <LTH> The Edg
183 4433 as won almost all their encounters to date.<t> Short emerged from the candidate
124 4421 ver. <CQ0> <t> Leap-horn gave him the date of the death of Pointed Shoes.<t> <C
153 4412 adequate nuclear weapons, kept up to date and based forward in Europe, our def
169 4404 financial climate in the country. To date I have only received five applicatio
155 4400 u work in the city?" and so on.<t> To date no machine has successfully fooled a
 56 4392 greement with the Chinese on a formal date for the resumption of diplomatic tie
185 4349 sible # says Bremner. The evidence to date, he says,suggests that men given a c
179 4320 ents. In one of the few such moves to date, KGF recently moved the management o
130 4308 t Tunisia has implicitly accepted the date of the 27th, so that would suggest t
161 4307 In New Cross # their greatest hit to date, is nowhere to be seen; but they do
156 4299 to enable the returner to keep up to date with developments.<t> Various other
141 4259 intervention until September 20, the date of the referendum, if necessary, and
143 4242 is expected to confirm April 9 as the date of the election. <h> Tories pin elec
188 4211 d, and though this is the first up-to date survey of its politics, it does not
 41 4205 MX.<M0X> Er communications. The copy date for the next issue of Foreword is th
159 4187 so they can keep their members up to date with what is happening in the indust
136 4077 ou post your order and payment by the date on the enclosed form. <LTH> <FCH> Bu
180 4071 s Milosevic's only live appearance to date came on one of its interview shows #
186 4050 3 months; none has become infected to date. In more than 70 incidents worldwide
178 4032 we are encouraged by our progress to date." In New York, John S. Reidy, analys
176 4019 now available but ask somebody up-to date.<M01> Mm.<F01> And of course compute
157 3999 ght of her man # <S0> Very much up to date, only been in service with our own f
129 3963 ard, Leo asked her for a date,and the date led to this. This deal has to be cas
 68 3950 nd its disposition, and of the likely date when the accumulated treasure, with
168 3937 be different. <LTH> As we come up to date, people # do it'' to be the same. Lo
172 3868 rhythm, 'Love # is his biggest hit to date. Chang's brand of lyrics label him a
 93 3843 he next three weeks # The NBL cut-off date for the finalisation of imports is n
163 3820 teach has not been very productive to date, nor is it likely to become more so
184 3774 n though the home loan was paid up to date. Few would ever have imagined they c
164 3770 ile, the multi-national menu is up-to date without being trendy: strikingly fre
165 3768 H> Sir-We are almost there, having to date raised <KPD> 92,000 in aid of lifebo
171 3767 nt clubs, this is Brainiak's story to date # B Sides'' features three God-bless
187 3717 s, and to considerable depths, but to date no detailed studies have been made o
162 3710 heir good work. Their achievements to date are quite amazing: land reclaimed; g
170 3699 the most impassioned Vedder vocal to date. He creates an opening mood of lonel
174 3679 ndia where we've got reasonably up to date statistics on population. Er there's
167 3679 ee billion bases, or coding units. To date, fossilised DNA has been extracted f
 98 3661 u do you have a sort of a prospective date for having the whole thing up and ru
158 3635 es with all the service records up to date. Abandoned only because arthritis ha
160 3609 tributed just less than dollars 3m to date. Most big state campaigns cost about
166 3603 notable two year-old performances to date # writes Dean Bailey <LTH> RESPONDIN
142 3595 Thatcher acted, bringing forward the date for a possible leadership election i

100 3582 nd FX that therefore the official pub date for the U K <F0X> <ZGY> <F0X> would
35 3560 Street, London ec88 2NG. <t> Closing date for the contest is January 7, 1992.
9 3533 they flew to Fort Worth to perform a date at the Dallas Hilton.Tina was wearin
140 3505 n the--on the--petrified tree and the date # And in all of my trips out to Mont
132 3409 ing my face again when he came to the date of birth, turning to the back to see
145 3394 dy else appears to have forgotten the date. Others feel the need to discuss the
38 3358 tel (about $1.2 million # The closing date for inclusion of properties is July
116 3323 last night to set 2000 as the target date for stabilising emissions of carbon
11 3311 her husband like to sometimes go on a date and spend the night in a hotel.Mrs.
59 3268 ishers, July 26 is the most important date in the year. It marks the anniversar
27 3246 ct entries selected after the closing date of Tuesday, August 10 will win the n
125 3211 was a model patient, remembering the date of every appointment and following a
50 3181 s. <LTH> Make sure you know the final date for accepting a place. Decline unwan
126 3175 e from the trial judge announcing the date of his execution in six weeks and on
101 3095 These measurements also give a quick date for that segment of the whole ice co
127 3090 s <CES> <t> So it would seem from the date of his birth # <t> My God # <t> He's
62 3067 960, pp. 181-8).An early inauguration date for the material product concept is
20 3043 however, with the addition of another date at the London Camden Falcon on Septe
139 2959 d war against Iraq # He also said the date is imminent and that a ground war ca
150 2956 endence, and Chart 91 is set for this date for Helsinki, the capital, for 12.00
135 2928 n the exact calendar months after the date the loan was opened. <LTH> Written q
51 2928 k and Barbara Sinatra after the final date of Mr Sinatra's London season. Today
131 2885 ccompanied by a printed report of the date, time and number of the attempted co
104 2880 res of Winners on June 20, the record date for Friday's special meeting to cons
103 2841 rew it and began again, with a record date of Aug. 29. Amdura has challenged th
146 2823 ssembly. The elections, such as their date and the voters' roll and even that m
86 2804 9) about British tennis are as out of date as the Dunlop Maxply in the attic. <
138 2796 about 13 # I can't remember what the date on that is--about 1773 or so # She--
133 2740 like the driver of a Hansom cab. The date is 29 March 1920.<t> <FCH> Above lef
6 2740 d declared their willingness to set a date for starting stage two of economic a
128 2686 ork on a regular basis and is not the date his family or household goods and ef
137 2679 t and your order # with a note of the date you sent it. Don't forget to give yo
149 2655 London,1981). Berlitz associated this date with the dire predictions given in G
108 2651 the 11th hour cancellation of Suede's date at the venue, and the closure of the
52 2645 Dublin on Saturday should set a firm date for an inter-governmental conference
117 2598 oard and set an implementation target date of January 1. <t> The working party
144 2592 debate the issue.<t> But setting the date is seen as little more than a pallia
123 2533 ded Tuesday to have a meeting on that date, the judge ordered the meeting held
7 2484 o longer enjoys the preparation for a date. 'Getting ready is part of the fun o
10 2482 Almost half thought she should set a date for stepping down; 35 per cent that
102 2460 the 'artwork' for him at the Reading date), going on to waltz until dawn with
75 2447 inistry denies there is a hold-up, no date has been set for a new round of talk
1 2428 LTH> Opening up for Metallica on a 65 date US tour, The Cult banged on with the
30 2412 rawn will each receive a kit. Closing date for entries is August 14. Standard r
28 2404 he first name drawn after the closing date on October 22, will receive a free L
32 2347 as soon as possible after the closing date. <LTH> 9 Send entries to: AP/Image H
111 2330 ticle ('Crime made easy') of the same date seems to have that problem. Can you
36 2267 ulled from our postbag on the closing date. <t> Over-18's only. News Internatio
29 2256 d Street, London se1 9LS. The closing date is Tuesday August 31, 1993 and the e
134 2195 rmehrung beim Umbau," which bears the date December 13, 1932 at the end.39 Ibid
147 2194 w. Leah claimed she knew by the third date that she wanted to become Mrs Winter
17 2149 e given here is set for this time and date, and for the capital, Paramaribo.<t>
12 2102 t of water # At one point he forgot a date that was sort of a simple date on wh
54 2062 > Do you play it cool after the first date? <LTH> Sarah If it was left that we

57 2036 iods of deep loneliness and grief.Her date of birth has been placed somewhere a
94 2011 lained that Iraq is offering only one date for a meeting, while he has offered
33 2004 to reach us no later than the closing date, July 31, 1993. <LTH> 1. TOGECAT <LT
15 1992 id Hogan, SM, has been adjourned to a date to be fixed.<dt> 930414 </dt> Cairns
37 1987 an 1500 immediately after the closing date # DAMIEN MARSH .. hard work has paid
21 1976 tion ready for critics of the bizarre date-rape story, 'What Actually Happened
13 1960 and flour when he stood her up for a date. But what has been the nature of the
194 1958 occur in the group of kouroi that we date the earliest.<t> JAFFE: The torso, o
67 1957 ople will return to church at a later date.I would like to invite everyone to a
151 1948 Edition # I'm Neal Conan.<t> On this date in 1956 the Republican Party nominat
199 1947 r after ordering the reactors # which date back to the 1950s # to be shut. They
69 1928 lf grows stronger, even if the likely date seems to recede towards the edge of
14 1925 to America for the Brando film and a date she wants to keep with Michelle Pfei
5 1925 Schedule </h> Playing this weekend:A Date With Judy (1948) Jane Powell plays a
47 1910 tion figure said the distant election date will give the ruling family time to
49 1869 -partisan. In announcing the election date, President Roh Tae Woo said there wa
91 1814 a puritan streak, and the concept of date" or 'acquaintance" rape reveals just
48 1776 of parties in parties in the Election date election parliament Albania Mar 23rd
99 1770 he Autumn Triangle. <LTH> Provisional date and venue for National Council 1992:
113 1754 osed October the 30th as the starting date # Mrs # Mandela's lawyer argued succ
8 1753 ance Group, who will present We Got A Date, Can't Take Johnny To The Funeral an
65 1681 t the same operation again at a later date. He may even, some analysts say,risk
81 1646 al government were swiftly put out of date yesterday by the President of Kazakh
66 1638 d by September 1993 and at some later date the US authorities will declare the
119 1614 hatching of meadow birds. After that date the mechanical cultivation of fields
53 1580 d--don't pick Red Lobster for a first date. Great lunch deals. Hours: Mon.-Thur
89 1558 ng data from instruments years out of date. Small craft allow the use of up-to-
55 1428 s had slept with a man on their first date and 39 per cent admitted to being un
118 1411 ns has two weeks from his termination date to appeal the decision. As for Randa
44 1393 he new Germany has chosen a different date # October 3 # Reunification Day to b
77 1372 ention a place. And so far there's no date fixed for the meeting. Until that ha
197 1365 It's not stated clearly back to what date this is effective # The decrees come
193 1347 the announced 20 March maiden voyage date. On 10 October the company released
4 1341 nue to try to get you into bed. <LTH> Date rape is at the forefront of all our
198 1266 we started restoring our murals which date back to the Portuguese era in the 14
97 1253 page May 3). Federal credit programs date back to the New Deal, and were meant
79 1223 t letting us go to work.<t> Dugan: No date has been set for the resumption of c
82 1181 ng them, is discriminating and out of date?They have had some support from lead
85 1138 F0> a newspaper it is slightly out of date but erm <F0X> Anything at all it'll
120 1137 wever, always say that he shares that date with my wife. <LTH> Dr L Keith Franc
96 1134 erly revised at the earliest possible date. <LTH> It is also unfortunate that a
42 1116 tland Yard says the children's deaths date back to 1984. Reports suggest that b
80 1106 others.<t> 1 Indefinite exclusion: no date is fixed for a return. Consideration
90 1104 ed administration,traditional, out-of date, a group of elderly men smoking ciga
34 1051 Kings Road, London sw10 OTE. Closing date is August 13. Normal rules apply. <h
84 1025 ticisms that their magazine is out of date or has lost its edge # Editor Zanne
107 1021 n vaults and galleries. Those in Rome date mainly from the third and early four
83 979 the S.E.2000 would be already out of date even before it first flew, and a new
39 933 call 008 812 772 for details.CLOSING DATE: August 13.DRAWN: August 20.MOBILE P
71 910 ogressive-punks play a one-off London date with Poisoned Electrick Head at New
76 879 population is a minority in Serbia.No date has been set for elections and there
95 867 xhaustion of a new mother. An opening date in June would have given her two pre
121 850 e Council by 1 April 1993. After that date, it will be an offence to run an unr
87 848 formation was always six hours out of date. I get an update from the senior for

```
 31  803 nner to be notified by phone. Closing date? box 29661 <LTH> CZECH 42, passive,
 88  791 l. Just as computers overwrite out-of date files on their disks, monks used to
 78  788 e, no proof, no dossier, no names, no date, no body.And as happens in all hosta
122  784 he first use of their cards from that date. They have until March 1, 1993, to c
 25  767 words and so we will see that sell-by date is no longer associated with perisha
 18  749 he next regularly slated announcement date is Sept. 18. A brand new directorate
 45  725 y 8. You may get a slightly different date by this short cut method than by add
109  714 ally signed and predated with today's date), his eyeglasses, a Koran, a Bible.
110  651 our chance of life was someone else's date with </h> death?;Steve Hyett;Part 2
 26  645 arded for decades, only an expert can date a garment. When a skirt length chang
190  641 here tier upon tier of rock-cut tombs date from as long ago as the thirteenth c
 60  635 as sensible of the priority of one in date. It was AD 450, that they beat the S
200  633 lossomed in the presence of women who date act ors and princes, dine in Milan a
196  621 ing to block a money-spinning Wembley date.Edwards hopes to convince FA Cup sem
 58  620  that it is vital.<p> Dr Salk's ideas date back a long way, but he has linked t
114  606  a visit with my son on such and such date else I would have been there. Probat
191  582 t of ownership.<t> Levinson: No trial date has been set yet for the Janis lawsu
 19  573 . <LTH> JULIAN COPE has added another date to his ''Head On'" tour. It's at Bra
 40  550 opening rounds. A mutually convenient date should then be set and green fees sh
115  525 le, the village is surrounded by tall date palms and lush green farmland. Its n
 92  511 ers announced the April 5th blast-off date following a flight review at the Ken
106  496 0> <t> Mark Keenan, 28, whose release date had been delayed by 28 days, was fou
 23  475 ed banner headlines about his # blind date'' escapade. <t> It was that sense of
 43  472 ears ahead of its intended deployment date. <t> Only ten navigators had been tr
105  465 0> Twenty years on from their release date, two albums look set to make this mo
 64  454 r, Sir Andrew Lloyd Webber. <t> Janet Date, a guide and former actress, is in h
 16  450 as been lined up as his dating agency date unbeknown to wife Alison Steadman wh
 72  437 10 cent a share # There is no meeting date set as yet.Pacarc said the issue to
 22  375 eels, and as an adult his first blind date.<t> Unidentified Woman (From Radio A
 24  351 LISTINGS <LTH> Concerts are listed by date, then by city # Classical Listings c
 70  320 ks (S) on design problems (location # date).Contract for Snabl already done, ps
192  312 ane 2222 No built <FCH> Aircraft Type Date Purpose of Design No built <FCH> Ans
 61  269 [heb.] shows that the poem is late in date. However,Phoenician inscriptions ear
 73  243 s portrayed.'' <t> <h> Benetton's new date;Motor Racing </h> <dt> 25 August 199
148  230 e asks you out again despite no third date action), you know you've built a fou
112  219 follow their previously announced six date tour, are priced <KPD> 8.50. Fans wi
195  177 inals in Filderstadt. <t> <h> Wembley date;Rugby League </h> <dt> 15 October 19
 46  139 egin to accumulate with each dividend date. drps really do serve an important f
 63  123 ER: Sue Waldram PRODUCER: Ferri Jahed DATE REC: 10 July 1990 TAPE NO: 90r/32k/0
  2   74 d </pres> <prod> Francis Mead </prod> Date Rec: 16 October 1990 Prog No: 90r/32
  3   50 details of each match you play. <LTH> DATE COMPETITION OPPONENT VENUE SCORE RES
```

e) Best-match Concordance Line Selection for Pearson Correlation 2 vs Representative: arts-prons/3/open/raw

157 5306 ght of her man # <SO> Very much up to da
163 4708 teach has not been very productive to da
175 4703 ties in which you've been involved to da
93 4187 he next three weeks # The NBL cut-off da
79 3947 t letting us go to work.<t> Dugan: No da
75 3806 inistry denies there is a hold-up, no da
152 3776 tary of state, brought Franklin up to da
58 3679 that it is vital.<p> Dr Salk's ideas da
140 3565 n the--on the--petrified tree and the da
189 3045 akes it the No. 1 film of the year to da
74 3006 to finish the record, before the next da
119 2970 hatching of meadow birds. After that da
122 2966 he first use of their cards from that da
159 2790 so they can keep their members up to da
143 2752 is expected to confirm April 9 as the da
188 2712 d, and though this is the first up-to da
138 2665 about 13 # I can't remember what the da
69 2530 lf grows stronger, even if the likely da
13 2512 and flour when he stood her up for a da
165 2498 H> Sir-We are almost there, having to da
4 2496 nue to try to get you into bed. <LTH> Da
161 2485 In New Cross # their greatest hit to da
168 2476 be different. <LTH> As we come up to da
25 2450 words and so we will see that sell-by da
176 2410 now available but ask somebody up-to da
111 2392 ticle ('Crime made easy') of the same da
155 2254 u work in the city?" and so on.<t> To da
6 2225 d declared their willingness to set a da
105 2219 O> Twenty years on from their release da
104 2201 res of Winners on June 20, the record da
88 2192 l. Just as computers overwrite out-of da
182 2144 iminalising of breaches in the law to da
80 2129 others.<t> l Indefinite exclusion: no da
55 2078 s had slept with a man on their first da
194 2072 occur in the group of kouroi that we da
116 2059 last night to set 2000 as the target da
84 2010 ticisms that their magazine is out of da
114 1999 a visit with my son on such and such da
76 1968 population is a minority in Serbia.No da
121 1934 e Council by 1 April 1993. After that da
60 1921 as sensible of the priority of one in da
83 1913 the S.E.2000 would be already out of da
177 1907 ning my 'firm grasp of the most up-to da
167 1867 ee billion bases, or coding units. To da
179 1853 ents. In one of the few such moves to da
102 1844 the 'artwork' for him at the Reading da
180 1843 s Milosevic's only live appearance to da
178 1828 we are encouraged by our progress to da
89 1826 ng data from instruments years out of da
96 1813 erly revised at the earliest possible da
72 1780 10 cent a share # There is no meeting da

174 1778 ndia where we've got reasonably up to date statistics on population. Er there's
187 1767 s, and to considerable depths, but to date no detailed studies have been made o
86 1736 9) about British tennis are as out of date as the Dunlop Maxply in the attic. <
153 1706 adequate nuclear weapons, kept up to date and based forward in Europe, our def
150 1646 endence, and Chart 91 is set for this date for Helsinki, the capital, for 12.00
156 1629 to enable the returner to keep up to date with developments.<t> Various other
15 1588 id Hogan, SM, has been adjourned to a date to be fixed.<dt> 930414 </dt> Cairns
181 1529 ll potential has not been realised to date owing to the ground.<t> In the Templ
186 1526 3 months; none has become infected to date. In more than 70 incidents worldwide
197 1479 It's not stated clearly back to what date this is effective # The decrees come
41 1473 MX.<M0X> Er communications. The copy date for the next issue of Foreword is th
172 1465 rhythm, 'Love # is his biggest hit to date. Chang's brand of lyrics label him a
66 1460 d by September 1993 and at some later date the US authorities will declare the
12 1456 t of water # At one point he forgot a date that was sort of a simple date on wh
123 1379 ded Tuesday to have a meeting on that date, the judge ordered the meeting held
61 1370 [heb.] shows that the poem is late in date. However,Phoenician inscriptions ear
154 1363 he said he would bring Mr Bush up to date on the issue:If we were forced to re
82 1363 ng them, is discriminating and out of date?They have had some support from lead
65 1313 t the same operation again at a later date. He may even, some analysts say,risk
118 1287 ns has two weeks from his termination date to appeal the decision. As for Randa
129 1278 ard, Leo asked her for a date,and the date led to this. This deal has to be cas
85 1272 F0> a newspaper it is slightly out of date but erm <F0X> Anything at all it'll
50 1266 s. <LTH> Make sure you know the final date for accepting a place. Decline unwan
130 1246 t Tunisia has implicitly accepted the date of the 27th, so that would suggest t
191 1244 t of ownership.<t> Levinson: No trial date has been set yet for the Janis lawsu
112 1241 follow their previously announced six date tour, are priced <KPD> 8.50. Fans wi
67 1232 ople will return to church at a later date.I would like to invite everyone to a
139 1222 d war against Iraq # He also said the date is imminent and that a ground war ca
59 1221 ishers, July 26 is the most important date in the year. It marks the anniversar
49 1220 -partisan. In announcing the election date, President Roh Tae Woo said there wa
146 1189 ssembly. The elections, such as their date and the voters' roll and even that m
184 1174 n though the home loan was paid up to date. Few would ever have imagined they c
164 1163 ile, the multi-national menu is up-to date without being trendy: strikingly fre
144 1132 debate the issue.<t> But setting the date is seen as little more than a pallia
145 1125 dy else appears to have forgotten the date. Others feel the need to discuss the
137 1099 t and your order # with a note of the date you sent it. Don't forget to give yo
43 1099 ears ahead of its intended deployment date. <t> Only ten navigators had been tr
35 1059 Street, London ec88 2NG. <t> Closing date for the contest is January 7, 1992.
32 1031 as soon as possible after the closing date. <LTH> 9 Send entries to: AP/Image H
158 1023 es with all the service records up to date. Abandoned only because arthritis ha
19 1003 . <LTH> JULIAN COPE has added another date to his ''Head On'' tour. It's at Bra
11 962 her husband like to sometimes go on a date and spend the night in a hotel.Mrs.
31 948 nner to be notified by phone. Closing date? box 29661 <LTH> CZECH 42, passive,
16 945 as been lined up as his dating agency date unbeknown to wife Alison Steadman wh
28 924 he first name drawn after the closing date on October 22, will receive a free L
128 908 ork on a regular basis and is not the date his family or household goods and ef
162 905 heir good work. Their achievements to date are quite amazing: land reclaimed; g
87 881 formation was always six hours out of date. I get an update from the senior for
169 870 financial climate in the country. To date I have only received five applicatio
9 867 they flew to Fort Worth to perform a date at the Dallas Hilton.Tina was wearin
94 856 lained that Iraq is offering only one date for a meeting, while he has offered
98 846 u do you have a sort of a prospective date for having the whole thing up and ru
199 833 r after ordering the reactors # which date back to the 1950s # to be shut. They
38 828 tel (about $1.2 million # The closing date for inclusion of properties is July

```
110  820 our chance of life was someone else's date with </h> death?;Steve Hyett;Part 2
 90  810 ed administration,traditional, out-of date, a group of elderly men smoking ciga
147  792 w. Leah claimed she knew by the third date that she wanted to become Mrs Winter
 56  784 greement with the Chinese on a formal date for the resumption of diplomatic tie
 20  784 however, with the addition of another date at the London Camden Falcon on Septe
142  779  Thatcher acted, bringing forward the date for a possible leadership election i
183  778 as won almost all their encounters to date.<t> Short emerged from the candidate
 54  763 > Do you play it cool after the first date? <LTH> Sarah If it was left that we
  7  735 o longer enjoys the preparation for a date. 'Getting ready is part of the fun o
127  732 s <CES> <t> So it would seem from the date of his birth # <t> My God # <t> He's
173  706 eally because er it's just been up-to date and it <M01> Mhm.<M02> I mean that's
 97  700  page May 3). Federal credit programs date back to the New Deal, and were meant
132  678 ing my face again when he came to the date of birth, turning to the back to see
 77  677 ention a place. And so far there's no date fixed for the meeting. Until that ha
  1  676 LTH> Opening up for Metallica on a 65 date US tour, The Cult banged on with the
166  669  notable two year-old performances to date # writes Dean Bailey <LTH> RESPONDIN
101  655  These measurements also give a quick date for that segment of the whole ice co
 95  647 xhaustion of a new mother. An opening date in June would have given her two pre
 14  646  to America for the Brando film and a date she wants to keep with Michelle Pfei
 57  609 iods of deep loneliness and grief.Her date of birth has been placed somewhere a
 52  609  Dublin on Saturday should set a firm date for an inter-governmental conference
 30  596 rawn will each receive a kit. Closing date for entries is August 14. Standard r
148  582 e asks you out again despite no third date action), you know you've built a fou
200  581 lossomed in the presence of women who date act ors and princes, dine in Milan a
100  567 nd FX that therefore the official pub date for the U K <F0X> <ZGY> <F0X> would
190  519 here tier upon tier of rock-cut tombs date from as long ago as the thirteenth c
126  512 e from the trial judge announcing the date of his execution in six weeks and on
 27  509 ct entries selected after the closing date of Tuesday, August 10 will win the n
 71  498 ogressive-punks play a one-off London date with Poisoned Electrick Head at New
160  462 tributed just less than dollars 3m to date. Most big state campaigns cost about
 44  446 he new Germany has chosen a different date # October 3 # Reunification Day to b
 78  431 e, no proof, no dossier, no names, no date, no body.And as happens in all hosta
 51  431 k and Barbara Sinatra after the final date of Mr Sinatra's London season. Today
117  404 oard and set an implementation target date of January 1. <t> The working party
106  402 0> <t> Mark Keenan, 28, whose release date had been delayed by 28 days, was fou
 23  378 ed banner headlines about his # blind date'' escapade. <t> It was that sense of
185  337 sible # says Bremner. The evidence to date, he says,suggests that men given a c
 45  324 y 8. You may get a slightly different date by this short cut method than by add
170  315  the most impassioned Vedder vocal to date. He creates an opening mood of lonel
151  307  Edition # I'm Neal Conan.<t> On this date in 1956 the Republican Party nominat
135  300 n the exact calendar months after the date the loan was opened. <LTH> Written q
 81  294 al government were swiftly put out of date yesterday by the President of Kazakh
 17  285 e given here is set for this time and date, and for the capital, Paramaribo.<t>
 26  281 arded for decades, only an expert can date a garment. When a skirt length chang
136  280 ou post your order and payment by the date on the enclosed form. <LTH> <FCH> Bu
 68  280 nd its disposition, and of the likely date when the accumulated treasure, with
 46  267 egin to accumulate with each dividend date. drps really do serve an important f
 99  261 he Autumn Triangle. <LTH> Provisional date and venue for National Council 1992:
193  258  the announced 20 March maiden voyage date. On 10 October the company released
171  257 nt clubs, this is Brainiak's story to date # B Sides'' features three God-bless
 39  255  call 008 812 772 for details.CLOSING DATE: August 13.DRAWN: August 20.MOBILE P
 47  237 tion figure said the distant election date will give the ruling family time to
192  233 ane 2222 No built <FCH> Aircraft Type Date Purpose of Design No built <FCH> Ans
198  218 we started restoring our murals which date back to the Portuguese era in the 14
```

```
124  203 ver. <CQ0> <t> Leap-horn gave him the date of the death of Pointed Shoes.<t> <C
 10  197  Almost half thought she should set a date for stepping down; 35 per cent that
 33  190 to reach us no later than the closing date, July 31, 1993. <LTH> 1. TOGECAT <LT
131  174 ccompanied by a printed report of the date, time and number of the attempted co
 91  168  a puritan streak, and the concept of date" or 'acquaintance" rape reveals just
 42  167 tland Yard says the children's deaths date back to 1984. Reports suggest that b
103  166 rew it and began again, with a record date of Aug. 29. Amdura has challenged th
 34  158  Kings Road, London sw10 OTE. Closing date is August 13. Normal rules apply. <h
 40  156 opening rounds. A mutually convenient date should then be set and green fees sh
 29  155 d Street, London se1 9LS. The closing date is Tuesday August 31, 1993 and the e
 70  142 ks (S) on design problems (location # date).Contract for Snabl already done, ps
113  139 osed October the 30th as the starting date # Mrs # Mandela's lawyer argued succ
 37  127 an 1500 immediately after the closing date # DAMIEN MARSH .. hard work has paid
  8  126 ance Group, who will present We Got A Date, Can't Take Johnny To The Funeral an
 62  118 960, pp. 181-8).An early inauguration date for the material product concept is
120  104 wever, always say that he shares that date with my wife. <LTH> Dr L Keith Franc
 36  100 ulled from our postbag on the closing date. <t> Over-18's only. News Internatio
133   86  like the driver of a Hansom cab. The date is 29 March 1920.<t> <FCH> Above lef
 21   83 tion ready for critics of the bizarre date-rape story, 'What Actually Happened
 48   82 of parties in parties in the Election date election parliament Albania Mar 23rd
141   75  intervention until September 20, the date of the referendum, if necessary, and
115   73 le, the village is surrounded by tall date palms and lush green farmland. Its n
196   71 ing to block a money-spinning Wembley date.Edwards hopes to convince FA Cup sem
125   68  was a model patient, remembering the date of every appointment and following a
 63   66 ER: Sue Waldram PRODUCER: Ferri Jahed DATE REC: 10 July 1990 TAPE NO: 90r/32k/0
  3   47 details of each match you play. <LTH> DATE COMPETITION OPPONENT VENUE SCORE RES
 18   46 he next regularly slated announcement date is Sept. 18. A brand new directorate
108   45 the 11th hour cancellation of Suede's date at the venue, and the closure of the
  2   39 d </pres> <prod> Francis Mead </prod> Date Rec: 16 October 1990 Prog No: 90r/32
107   38 n vaults and galleries. Those in Rome date mainly from the third and early four
109   37 ally signed and predated with today's date), his eyeglasses, a Koran, a Bible.
 22   34 eels, and as an adult his first blind date.<t> Unidentified Woman (From Radio A
 64   33 r, Sir Andrew Lloyd Webber. <t> Janet Date, a guide and former actress, is in h
149   28 London,1981). Berlitz associated this date with the dire predictions given in G
 92   28 ers announced the April 5th blast-off date following a flight review at the Ken
 53   23 d--don't pick Red Lobster for a first date. Great lunch deals. Hours: Mon.-Thur
 73   21 s portrayed.'' <t> <h> Benetton's new date;Motor Racing </h> <dt> 25 August 199
 24    8 LISTINGS <LTH> Concerts are listed by date, then by city # Classical Listings c
195    4 inals in Filderstadt. <t> <h> Wembley date;Rugby League </h> <dt> 15 October 19
134    3 rmehrung beim Umbau," which bears the date December 13, 1932 at the end.39 Ibid
  5    2  Schedule </h> Playing this weekend:A Date With Judy (1948) Jane Powell plays a
```

f) Best-match Concordance Line Selection for Pearson Correlation 2 vs Usable: arts-prons/2/fixed/rel

188 1887 d, and though this is the first up-to date survey of its politics, it does not
157 1853 ght of her man # <SO> Very much up to date, only been in service with our own f
159 1840 so they can keep their members up to date with what is happening in the indust
176 1831 now available but ask somebody up-to date.<M01> Mm.<F01> And of course compute
152 1761 tary of state, brought Franklin up to date on the bloodshed in his beloved Fran
168 1703 be different. <LTH> As we come up to date, people # do it'' to be the same. Lo
184 1677 n though the home loan was paid up to date. Few would ever have imagined they c
153 1632 adequate nuclear weapons, kept up to date and based forward in Europe, our def
177 1598 ning my 'firm grasp of the most up-to date trauma procedures". <t> The referenc
154 1595 he said he would bring Mr Bush up to date on the issue:If we were forced to re
174 1586 ndia where we've got reasonably up to date statistics on population. Er there's
156 1565 to enable the returner to keep up to date with developments.<t> Various other
158 1515 es with all the service records up to date. Abandoned only because arthritis ha
173 1512 eally because er it's just been up-to date and it <M01> Mhm.<M02> I mean that's
164 1506 ile, the multi-national menu is up-to date without being trendy: strikingly fre
 79 1223 t letting us go to work.<t> Dugan: No date has been set for the resumption of c
132 1049 ing my face again when he came to the date of birth, turning to the back to see
180 925 s Milosevic's only live appearance to date came on one of its interview shows #
161 851 In New Cross # their greatest hit to date, is nowhere to be seen; but they do
 90 749 ed administration,traditional, out-of date, a group of elderly men smoking ciga
 84 683 ticisms that their magazine is out of date or has lost its edge # Editor Zanne
 85 670 F0> a newspaper it is slightly out of date but erm <F0X> Anything at all it'll
 75 658 inistry denies there is a hold-up, no date has been set for a new round of talk
 82 656 ng them, is discriminating and out of date?They have had some support from lead
 88 653 1. Just as computers overwrite out-of date files on their disks, monks used to
 86 648 9) about British tennis are as out of date as the Dunlop Maxply in the attic. <
 83 631 the S.E.2000 would be already out of date even before it first flew, and a new
 76 631 population is a minority in Serbia.No date has been set for elections and there
175 628 ties in which you've been involved to date?<M01> Er the spectrum of that would
 81 624 al government were swiftly put out of date yesterday by the President of Kazakh
167 609 ee billion bases, or coding units. To date, fossilised DNA has been extracted f
 87 606 formation was always six hours out of date. I get an update from the senior for
 89 601 ng data from instruments years out of date. Small craft allow the use of up-to-
  6 582 d declared their willingness to set a date for starting stage two of economic a
172 575 rhythm, 'Love # is his biggest hit to date. Chang's brand of lyrics label him a
 15 571 id Hogan, SM, has been adjourned to a date to be fixed.<dt> 930414 </dt> Cairns
 57 557 iods of deep loneliness and grief.Her date of birth has been placed somewhere a
 13 507 and flour when he stood her up for a date. But what has been the nature of the
 38 475 tel (about $1.2 million # The closing date for inclusion of properties is July
186 457 3 months; none has become infected to date. In more than 70 incidents worldwide
191 448 t of ownership.<t> Levinson: No trial date has been set yet for the Janis lawsu
181 445 ll potential has not been realised to date owing to the ground.<t> In the Templ
178 432 we are encouraged by our progress to date." In New York, John S. Reidy, analys
155 429 u work in the city?" and so on.<t> To date no machine has successfully fooled a
116 416 last night to set 2000 as the target date for stabilising emissions of carbon
 35 414 Street, London ec88 2NG. <t> Closing date for the contest is January 7, 1992.
 30 407 rawn will each receive a kit. Closing date for entries is August 14. Standard r
150 396 endence, and Chart 91 is set for this date for Helsinki, the capital, for 12.00
 56 385 greement with the Chinese on a formal date for the resumption of diplomatic tie
 80 380 others.<t> 1 Indefinite exclusion: no date is fixed for a return. Consideration
 67 368 ople will return to church at a later date.I would like to invite everyone to a

66 367 d by September 1993 and at some later date the US authorities will declare the
27 364 ct entries selected after the closing date of Tuesday, August 10 will win the n
163 354 teach has not been very productive to date, nor is it likely to become more so
121 351 e Council by 1 April 1993. After that date, it will be an offence to run an unr
165 349 H> Sir-We are almost there, having to date raised <KPD> 92,000 in aid of lifebo
189 347 akes it the No. 1 film of the year to date and the biggest April release in the
127 337 s <CES> <t> So it would seem from the date of his birth # <t> My God # <t> He's
169 323 financial climate in the country. To date I have only received five applicatio
65 318 t the same operation again at a later date. He may even, some analysts say,risk
119 307 hatching of meadow birds. After that date the mechanical cultivation of fields
34 305 Kings Road, London sw10 OTE. Closing date is August 13. Normal rules apply. <h
187 304 s, and to considerable depths, but to date no detailed studies have been made o
101 296 These measurements also give a quick date for that segment of the whole ice co
29 294 d Street, London se1 9LS. The closing date is Tuesday August 31, 1993 and the e
28 292 he first name drawn after the closing date on October 22, will receive a free L
111 288 ticle ('Crime made easy') of the same date seems to have that problem. Can you
25 288 words and so we will see that sell-by date is no longer associated with perisha
197 280 It's not stated clearly back to what date this is effective # The decrees come
32 279 as soon as possible after the closing date. <LTH> 9 Send entries to: AP/Image H
77 277 ention a place. And so far there's no date fixed for the meeting. Until that ha
72 264 10 cent a share # There is no meeting date set as yet.Pacarc said the issue to
52 247 Dublin on Saturday should set a firm date for an inter-governmental conference
162 230 heir good work. Their achievements to date are quite amazing: land reclaimed; g
131 229 ccompanied by a printed report of the date, time and number of the attempted co
10 224 Almost half thought she should set a date for stepping down; 35 per cent that
93 217 he next three weeks # The NBL cut-off date for the finalisation of imports is n
183 210 as won almost all their encounters to date.<t> Short emerged from the candidate
94 189 lained that Iraq is offering only one date for a meeting, while he has offered
74 183 to finish the record, before the next date of the tour in Lisbon. <LTH> The Edg
33 179 to reach us no later than the closing date, July 31, 1993. <LTH> 1. TOGECAT <LT
185 175 sible # says Bremner. The evidence to date, he says,suggests that men given a c
55 168 s had slept with a man on their first date and 39 per cent admitted to being un
198 166 we started restoring our murals which date back to the Portuguese era in the 14
37 164 an 1500 immediately after the closing date # DAMIEN MARSH .. hard work has paid
97 163 page May 3). Federal credit programs date back to the New Deal, and were meant
42 163 tland Yard says the children's deaths date back to 1984. Reports suggest that b
199 156 r after ordering the reactors # which date back to the 1950s # to be shut. They
51 155 k and Barbara Sinatra after the final date of Mr Sinatra's London season. Today
171 146 nt clubs, this is Brainiak's story to date # B Sides'' features three God-bless
166 139 notable two year-old performances to date # writes Dean Bailey <LTH> RESPONDIN
117 136 oard and set an implementation target date of January 1. <t> The working party
102 133 the 'artwork' for him at the Reading date), going on to waltz until dawn with
151 131 Edition # I'm Neal Conan.<t> On this date in 1956 the Republican Party nominat
98 130 u do you have a sort of a prospective date for having the whole thing up and ru
17 127 e given here is set for this time and date, and for the capital, Paramaribo.<t>
138 123 about 13 # I can't remember what the date on that is--about 1773 or so # She--
12 123 t of water # At one point he forgot a date that was sort of a simple date on wh
70 118 ks (S) on design problems (location # date).Contract for Snabl already done, ps
145 110 dy else appears to have forgotten the date. Others feel the need to discuss the
143 108 is expected to confirm April 9 as the date of the election. <h> Tories pin elec
190 106 here tier upon tier of rock-cut tombs date from as long ago as the thirteenth c
59 105 ishers, July 26 is the most important date in the year. It marks the anniversar
39 101 call 008 812 772 for details.CLOSING DATE: August 13.DRAWN: August 20.MOBILE P
123 97 ded Tuesday to have a meeting on that date, the judge ordered the meeting held

96	97	erly revised at the earliest possible	date. <LTH> It is also unfortunate that a
182	93	iminalising of breaches in the law to	date.<t> The ruling Christian Democrats a
103	93	rew it and began again, with a record	date of Aug. 29. Amdura has challenged th
20	93	however, with the addition of another	date at the London Camden Falcon on Septe
60	92	as sensible of the priority of one in	date. It was AD 450, that they beat the S
9	90	they flew to Fort Worth to perform a	date at the Dallas Hilton.Tina was wearin
140	87	n the--on the--petrified tree and the	date # And in all of my trips out to Mont
105	87	0> Twenty years on from their release	date, two albums look set to make this mo
106	86	0> <t> Mark Keenan, 28, whose release	date had been delayed by 28 days, was fou
147	85	w. Leah claimed she knew by the third	date that she wanted to become Mrs Winter
122	84	he first use of their cards from that	date. They have until March 1, 1993, to c
104	80	res of Winners on June 20, the record	date for Friday's special meeting to cons
91	80	a puritan streak, and the concept of	date" or 'acquaintance" rape reveals just
40	78	opening rounds. A mutually convenient	date should then be set and green fees sh
126	76	e from the trial judge announcing the	date of his execution in six weeks and on
71	75	ogressive-punks play a one-off London	date with Poisoned Electrick Head at New
19	67	. <LTH> JULIAN COPE has added another	date to his ''Head On'" tour. It's at Bra
54	65	> Do you play it cool after the first	date? <LTH> Sarah If it was left that we
160	59	tributed just less than dollars 3m to	date. Most big state campaigns cost about
139	56	d war against Iraq # He also said the	date is imminent and that a ground war ca
95	56	xhaustion of a new mother. An opening	date in June would have given her two pre
45	56	y 8. You may get a slightly different	date by this short cut method than by add
41	53	MX.<M0X> Er communications. The copy	date for the next issue of Foreword is th
114	50	a visit with my son on such and such	date else I would have been there. Probat
43	50	ears ahead of its intended deployment	date. <t> Only ten navigators had been tr
23	50	ed banner headlines about his # blind	date'' escapade. <t> It was that sense of
192	46	ane 2222 No built <FCH> Aircraft Type	Date Purpose of Design No built <FCH> Ans
144	45	debate the issue.<t> But setting the	date is seen as little more than a pallia
21	45	tion ready for critics of the bizarre	date-rape story, 'What Actually Happened
124	43	ver. <CQ0> <t> Leap-horn gave him the	date of the death of Pointed Shoes.<t> <C
194	42	occur in the group of kouroi that we	date the earliest.<t> JAFFE: The torso, o
130	42	t Tunisia has implicitly accepted the	date of the 27th, so that would suggest t
141	41	intervention until September 20, the	date of the referendum, if necessary, and
53	41	d--don't pick Red Lobster for a first	date. Great lunch deals. Hours: Mon.-Thur
36	40	ulled from our postbag on the closing	date. <t> Over-18's only. News Internatio
112	39	follow their previously announced six	date tour, are priced <KPD> 8.50. Fans wi
108	38	the 11th hour cancellation of Suede's	date at the venue, and the closure of the
120	37	wever, always say that he shares that	date with my wife. <LTH> Dr L Keith Franc
63	32	ER: Sue Waldram PRODUCER: Ferri Jahed	DATE REC: 10 July 1990 TAPE NO: 90r/32k/0
7	32	o longer enjoys the preparation for a	date. 'Getting ready is part of the fun o
4	32	nue to try to get you into bed. <LTH>	Date rape is at the forefront of all our
136	30	ou post your order and payment by the	date on the enclosed form. <LTH> <FCH> Bu
135	30	n the exact calendar months after the	date the loan was opened. <LTH> Written q
118	30	ns has two weeks from his termination	date to appeal the decision. As for Randa
50	26	s. <LTH> Make sure you know the final	date for accepting a place. Decline unwan
193	25	the announced 20 March maiden voyage	date. On 10 October the company released
31	25	nner to be notified by phone. Closing	date? box 29661 <LTH> CZECH 42, passive,
128	24	ork on a regular basis and is not the	date his family or household goods and ef
107	24	n vaults and galleries. Those in Rome	date mainly from the third and early four
99	24	he Autumn Triangle. <LTH> Provisional	date and venue for National Council 1992:
1	24	LTH> Opening up for Metallica on a 65	date US tour, The Cult banged on with the
24	23	LISTINGS <LTH> Concerts are listed by	date, then by city # Classical Listings c
62	22	960, pp. 181-8).An early inauguration	date for the material product concept is
22	22	eels, and as an adult his first blind	date.<t> Unidentified Woman (From Radio A

```
  2   22 d </pres> <prod> Francis Mead </prod> Date Rec: 16 October 1990 Prog No: 90r/32
 68   21 nd its disposition, and of the likely date when the accumulated treasure, with
129   20 ard, Leo asked her for a date,and the date led to this. This deal has to be cas
 11   20 her husband like to sometimes go on a date and spend the night in a hotel.Mrs.
142   19  Thatcher acted, bringing forward the date for a possible leadership election i
110   19 our chance of life was someone else's date with </h> death?;Steve Hyett;Part 2
 69   19 lf grows stronger, even if the likely date seems to recede towards the edge of
146   18 ssembly. The elections, such as their date and the voters' roll and even that m
 78   17 e, no proof, no dossier, no names, no date, no body.And as happens in all hosta
 16   17 as been lined up as his dating agency date unbeknown to wife Alison Steadman wh
 18   14 he next regularly slated announcement date is Sept. 18. A brand new directorate
133   13  like the driver of a Hansom cab. The date is 29 March 1920.<t> <FCH> Above lef
  8   12 ance Group, who will present We Got A Date, Can't Take Johnny To The Funeral an
113   11 osed October the 30th as the starting date # Mrs # Mandela's lawyer argued succ
195   10 inals in Filderstadt. <t> <h> Wembley date;Rugby League </h> <dt> 15 October 19
137   10 t and your order # with a note of the date you sent it. Don't forget to give yo
125   10  was a model patient, remembering the date of every appointment and following a
 44   10 he new Germany has chosen a different date # October 3 # Reunification Day to b
 61    9 [heb.] shows that the poem is late in date. However,Phoenician inscriptions ear
 49    9 -partisan. In announcing the election date, President Roh Tae Woo said there wa
 47    9 tion figure said the distant election date will give the ruling family time to
 14    9  to America for the Brando film and a date she wants to keep with Michelle Pfei
115    8 le, the village is surrounded by tall date palms and lush green farmland. Its n
179    7 ents. In one of the few such moves to date, KGF recently moved the management o
100    6 nd FX that therefore the official pub date for the U K <FOX> <ZGY> <FOX> would
 48    6 of parties in parties in the Election date election parliament Albania Mar 23rd
148    5 e asks you out again despite no third date action), you know you've built a fou
 73    4 s portrayed.'' <t> <h> Benetton's new date;Motor Racing </h> <dt> 25 August 199
 26    4 arded for decades, only an expert can date a garment. When a skirt length chang
200    3 lossomed in the presence of women who date act ors and princes, dine in Milan a
 58    3  that it is vital.<p> Dr Salk's ideas date back a long way, but he has linked t
196    2 ing to block a money-spinning Wembley date.Edwards hopes to convince FA Cup sem
170    1  the most impassioned Vedder vocal to date. He creates an opening mood of lonel
134    1 rmehrung beim Umbau," which bears the date December 13, 1932 at the end.39 Ibid
 46    1 egin to accumulate with each dividend date. drps really do serve an important f
  5    1  Schedule </h> Playing this weekend:A Date With Judy (1948) Jane Powell plays a
  3    1 details of each match you play. <LTH> DATE COMPETITION OPPONENT VENUE SCORE RES
149    0 London,1981). Berlitz associated this date with the dire predictions given in G
109    0 ally signed and predated with today's date), his eyeglasses, a Koran, a Bible.
 92    0 ers announced the April 5th blast-off date following a flight review at the Ken
 64    0 r, Sir Andrew Lloyd Webber. <t> Janet Date, a guide and former actress, is in h
```