International Journal of Corpus Linguistics. (2002) 7, 2: 215-44.

Two quantitative methods of studying phraseology in English

Michael Stubbs University of Trier, Germany

Word frequency lists are a standard resource for many theoretical, descriptive and applied questions. However, due to severe problems of definition, there are no equivalent lists which give the frequency of phrases. This paper proposes two independent methods of studying the frequent phraseology of English. First, using a data-base of the most frequent collocations between word-forms in a 200-million word corpus, the strength of attraction between pairs of content words is discussed. Second, using a corpus of 2.5 million words, some of the most frequent phrases, in the sense of strings of uninterrupted word-forms, are identified, and their lexical, grammatical and semantic features are discussed.

Keywords: phraseology, collocations, word frequency, recurrent word-combinations

Introduction

Word frequency lists are a standard resource for linguists. However, although lists of well-known phrases are available in many taxonomies and dictionaries of collocations, only very limited frequency data are available. There are two obvious reasons for this lack of data on the frequency of phrases. First, although the phraseological nature of language has been thoroughly documented by corpus studies, there is still a tendency, following hundreds of years of lexicographic tradition, to think of individual words, rather than phrases, as the basic units of language. Second, since there are severe problems in defining phrasal units in corpora, it is difficult to know what to count. Indeed, it is doubtful if there could be a definitive 'phrase frequency list', since the units in question are so variable, and can be defined at such different levels of abstraction. Nevertheless, it is perfectly possible to investigate quantitative aspects of different kinds of multi-word units. This paper uses two different definitions of such units, identifies for each case some of the most frequent phrases in English, and discusses some characteristics of their constituent lexis, grammar and semantics.

A first method: collocations

The first part of the paper discusses briefly some general characteristics of collocations, and presents some results from a study of the extent and strength of collocations across the most frequent content words (that is, nouns, verbs, adjectives and adverbs) in English. Here, a quantitative approach to phraseology requires a large number of examples, which are drawn from a large corpus of data,

since only this allows generalizations to be made about collocational relations across the whole lexicon. In an ideal world, such generalizations would be based on a comprehensive list of all of the most frequent collocations, down to some frequency cut-off point. In the real world, resources are limited, and the present study uses a data-base (Cobuild 1995a) of around 200,000 collocations between pairs of frequent content words in English, and presents an analysis of a sub-sample of these pairs.

Terminology in this area is very variable, and I will use the following terms and definitions. I will use a purely statistical concept of collocation to refer to the habitual co-occurrence of words (examples involving both word-forms and lemmas are given). LEMMAS (in upper case) means a class of word-forms (in lower case italics). For example the lemma TAKE includes the word-forms *take, takes, taking, taken* and *took*. "Meanings" are in double quotes. I will talk of a node-word co-occurring with collocates (word-forms or lemmas) in a span of words to left and right. The data are from a span of 4:4. Collocates appear in diamond brackets: node <collocates>. A statement such as

– node 100 <collocate-1 10%, collocate-2, -3, -4 ...> 30%

is to be read as follows: a node-word occurs 100 times and co-occurs with a single collocate in 10 per cent of cases, and together with other collocates in a total of 30 per cent of cases. Since frequencies of occurrence depend on the size of the corpus, I often follow the convention of normalizing frequencies to their estimated occurrence in one million running words. All examples cited are attested in corpus data.

Collocations as idiosyncratic?

The phraseology of English certainly reveals many arbitrary constraints, and it is easy to find examples of collocation which make it look like an idiosyncratic and peripheral phenomenon. For example, one can say both *at a young age* and *at an old age*; but although one can say *in his old age*, one cannot say *in his young age*. In addition, there are words which occur only in fixed phrases and have no independent existence, such as *dint* (*by dint of*) and *sleight* (*sleight of hand*); and there are many verbs which usually co-occur with only one noun, such as *shrug one's shoulders* and *foot the bill*. In addition, collocations appear to vary idiosyncratically across languages:

– ask a question; poser une question; eine Frage stellen
– set a question (an exam question); ein Thema stellen
– make an application; einen Antrag stellen
– take (or make) a decision; eine Entscheidung treffen

– take (*make) a step forward; einen Schritt nach vorne machen

A common phenomenon is that one language uses a dummy (lexically empty) verb such as MAKE or FAIRE, in cases where another language uses a specific verb such as TREFFEN (= "meet" or "hit"). Even Melˇcuk (1996: 76), in his attempt at a comprehensive classification of the many general and abstract meanings expressed by word-combinations, talks of the 'basically idiosyncratic character' of these relations.

However, there have long been scholars who have emphasized the large number of predictable collocations and prefabricated expressions in everyday language use: Harold Palmer and J. R. Firth emphasized this point in the 1930s, and more recently, Hymes (1962: 126–127) talks of a 'vast portion of verbal behaviour' consisting of recurrent patterns, Bolinger (1976: 1) talks of 'an incredibly large number of prefabs', and Pawley and Syder (1983: 213) talk of 'several hundreds of thousands' of ready-made expressions and give an informal indication of how this number was estimated. Wray (2002: 7–11) provides a thorough discussion of these and similar ideas.

Much work on collocations is characterized by one of these two tendencies. There is valuable and detailed work, on individual words or small lexical sets (which sometimes over-emphasizes idiosyncratic cases), and there is work which emphasizes the frequency and centrality of the phenomenon (but often makes only vague quantitative statements). The ideal would be to combine the best of both approaches, so as to make more precise quantitative generalizations about collocations across the whole of the vocabulary of a language. This would require a method which meets certain criteria. Most importantly, the basic data must be a large and unbiased sample of collocations. Such a sample automatically gives priority to the most frequent words in their most frequent collocations, and therefore to central tendencies in the vocabulary, rather than to infrequent and idiosyncratic examples. In addition, frequency data also allow us to quantify the strength of attraction between node and collocates.

A long-term criterion for such work, which would take into account less frequent collocations, is comprehensive coverage of the vocabulary. As Miller (1998: xv) ironically notes, much work underestimates the importance of this criterion:

> An author might propose a semantic theory and illustrate it with 20 or 50 English words (usually nouns), leaving the other 100,000 words of English as an exercise for the reader.

Here, I have the more modest aim of illustrating a method and presenting findings from a sample of data, and the medium-term aim of stating the patterns which characterize the most frequent collocations across the most frequent vocabulary.

The example of IMPLEMENT (verb)

A first isolated example will illustrate the approach. In the corpus of 200-million running words used for the Cobuild (1995a) data-base, described below, the word-form *implemented* occurs over 1,900 times, with a very clear semantic preference: the most frequent noun collocates are words meaning "plan" and "change", and these words themselves occur as observable collocates.

– implemented <plan 5%, reform(s) policy/ies, measures, changes, programme, recommendations, resolutions, agreement, proposals, scheme> 24%

If we compare such data on a wide sample of words, we can then ask whether this case is typical. Does implemented exert a stronger collocational attraction on its surrounding collocates than average? The brief answer is that the strength of attraction between the node and its top collocate falls exactly within the norm for the vocabulary as a whole (see below). The attraction between the node and the whole set of words meaning approximately "plan" is probably rather stronger than average, though by no means extreme.

Data-base

A large sample of node-words and their collocates which can be used for quantitative study is available as Cobuild English Collocations on CD-ROM (Cobuild 1995a). This data-base was constructed as follows. From a 200-million word corpus, the 10,000 most frequent word-forms were extracted, and for each of these node head-words, the 20 most frequent collocates (down to a frequency cut-off of 15 across the whole corpus) were extracted in a span of 4:4. For each collocate, 20 concordance lines were extracted at random. (Only content words appear in the primary lists of head-words and collocates, and only content words are studied here. Grammatical words can be studied via supplementary lists.) The whole data-base therefore consists of around 4 million concordance lines, which each have a rough description of their text-type provenance, such as British fiction and American journalism. (To be strictly accurate, since a frequency cut-off point of 15 was set for node-collocate attraction, not all node- words are listed with a full 20 collocates, and there are somewhat fewer than 4 million lines.)

Such a data-base allows us to investigate: the most frequent collocations; the strength of attraction between node and collocate; the extent of variation in this attraction; and the semantic relations which recur most frequently between nodes and collocates.

A very simple question

One of the simplest questions is: What are the most frequent node-collocate pairs? Across the 10,000 pairs of node and top collocate (where both are content words), around 40 pairs occur more than 50 times per million words, e.g.:

– no longer; same time; last month; some people

And around 1,500 pairs occur more than 5 times per million words, either as uninterrupted phrases or collocate pairs separated by other words, e.g.:

– bad news; living room; spend time; take part
– dozen <half>; parents <children>; throat <clears>; value <money>

Native speakers recognize such collocations – in retrospect – as entirely banal, although they would be unable to retrieve them from introspection, other than on an individual and unsystematic basis. Fox (1987) reports a small experiment in which she tried to elicit the typical collocates of common words from native speakers. She concludes that, once they are told what the most frequent collocate is, 'it is so obvious that no-one can imagine not guessing it correctly', but 'the important thing is that they had not' (Fox 1987: 146). De Beaugrande (1999: 247) makes the same point that intuitions are only weakly predictive, but strongly 'retrodictive'.

Non-transparent (non-compositional) collocations

A second question is: How many frequent node-collocate pairs can be understood compositionally? For example, a frequent phrase in the data-base is *bad news*. If you know the meaning of the individual words, then you know the meaning of the phrase, since it is predictable by rule, as an intersection of the meanings of the constituent words. However, a phrase such as *high school* cannot be understood analogously to *high building*. There is a possible paradigmatic contrast with *low building*, but not with **low school*. Words can be inserted into the middle of one phrase, *high and elegant building*, but not the other, **high and elegant school*. And when translated morpheme-by-morpheme into other languages, the meaning alters. For example, German *Hochschule* means not "high school", but "university". It would probably not be possible to quantify exactly which collocations can and cannot be understood compositionally. After all, *high school* has something to do with a school which is high in some sense, but the meaning of the whole is more than the sum of the parts: it has both internal grammatical structure and also semantic unity. These general characteristics of collocations are discussed in many places, such as Wray (2002: 44–66).

Depending on how generous one's definition is, amongst the most frequent combinations of node and top collocate (> 5 times per million words), between 10 and 20 per cent are non-transparent, e.g.:

– box office; first leg; higher education; single currency; straight away; take part

Amongst the least frequent collocations in the data-base (< 0.3 per million), around 10 per cent are non-transparent, e.g.:

– emotional blackmail; forbidden fruit; heavenly bodies; snakes (and) ladders; title bout

Such figures probably underestimate cases of non-compositionality, since they are figures only for nodes and their single top collocates. For example, the collocates of *eye* indicate its use in several separate idiomatic expressions (cf. Sinclair 1991: 495):

– eye <keep, caught, public, blind, private>

A sub-set of non-compositional phrases are those that would often be called true idioms, since their meanings are hardly, if at all, predictable from the sum of their parts. Here are some examples, still of two word-forms, node and top collocate:

– axe <grind 7%>; chalk <cheese 4%>; feather <nest 8%>; thumb <rule 12%>; cheek <tongue 14%>; tongue <cheek 9%>

Such examples are not very frequent, as Moon (1998: 57–74) documents, but they are important, because they show that the most frequent uses of a frequent word may be more often idiomatic than literal, and have little or nothing to do with what seems to be its literal stand-alone meaning. Consider, for example, the collocates of *axe*.

– axe <grind 7%, FACE 7%, FALL 3%, WIELD 3%, plans, murderer, pick, cut, decision, battle>

The most frequent individual collocate is due to an idiom which occurs in slightly different forms (*with an axe to grind*; *have no axe to grind*; etc.). The most frequent uses, especially in journalism, are metaphorical (e.g. *jobs face the axe*; *plans to axe jobs*). Some collocates are due to literal uses (*axe murderer, pick-axe, battle-axe*), but some of these are also metaphorical (*the old battle-axe gave me a glare*). So, as with *eye*, there are several different idiomatic and non-literal uses, which are more frequent than the literal uses.

Attraction between single word-forms (node and top collocate)

I now start to make broader generalizations about the strength of attraction between pairs of individual word-forms. Over 20 per cent of the most frequent content words co-occur with one specific collocate in over 10 per cent of occurrences; over 65 per cent co-occur with one specific collocate in over 5 per cent of occurrences. Conversely, few words have less than one chance in 50 of co-occurring with one specific collocate; hardly any words in the data-base have a strength of attraction, between node and top collocate, of less than one per cent.

Statements in this form are perhaps rather difficult to follow: a certain percentage of node-words exerts a certain strength of attraction. However, they are easier to interpret with some examples, all still of just two word-forms, node and top collocate, as follows. Some node words have a strength of attraction of 30 per cent and over. This covers a wide range of attraction, but only a small number of nodes (ca 2 per cent) are involved. Examples are:

– fashioned <old 86%>; eighteenth <century 77%>; cloves <garlic 63%>; awaited <long 51%>; coronary <disease 43%>; basics <back 37%>

Still a small number (ca 3 per cent) have a strength of attraction of between 20 and 29 per cent. Examples are:

– profile <high 28%>; tricks <dirty 25%>

Rather more nodes (ca 18 per cent) have a strength of attraction of between 10 and 19 per cent. Examples are:

– bronze <medal 14%>; insufficient <evidence 11%>

The large majority of nodes (ca 74 per cent) have a strength of attraction of between 2 and 9 per cent. Around 44 per cent of nodes have a strength of attraction of between 5 and 9 per cent. Examples are:

– desert <island 6%>; await <outcome 5%>

And around 30 per cent of nodes have a strength of attraction of between 2 and 4 per cent. Examples are:

– expecting <baby 4%>; fit <keep 3%>

Perhaps most striking of all is that very few nodes (ca 1.5 per cent) have only a weak attraction of under 2 per cent. Amongst these rare examples are:

– ashes <phoenix>; castle <old>; renaissance <enjoying>

Very few nodes indeed (around one in a thousand) have a strength of attraction of under 1 per cent. One example is *victor <emerged 0.7%>*.

Figure 1 shows a graphic summary of these points. The graph shows data from a sample of 1,000 word-pairs: node and top collocate. The x-axis shows the strength of attraction within the pairs. This is here cut off at 25 per cent: there are node-words with a higher strength of attraction (examples were given above), but the graph has flattened out by this point, so I have represented only around 950 data points on the curve. The y-axis shows the number of node-words which have a given strength of attraction. There is little point in giving an average (mean) strength of attraction, since there is too much variation, but the mode is very clear: most node-words have a strength of attraction of between 2 and 10 per cent. Above 10 per cent, there are many fewer node-words, and the curve rapidly flattens out. What is perhaps most striking is that the graph falls off even more steeply in the other direction: very few node-words have only a weak strength of attraction, of less than 2 per cent.

The extent and strength of such collocational attraction is much greater than generally realised and rarely taken into account in linguistic description.

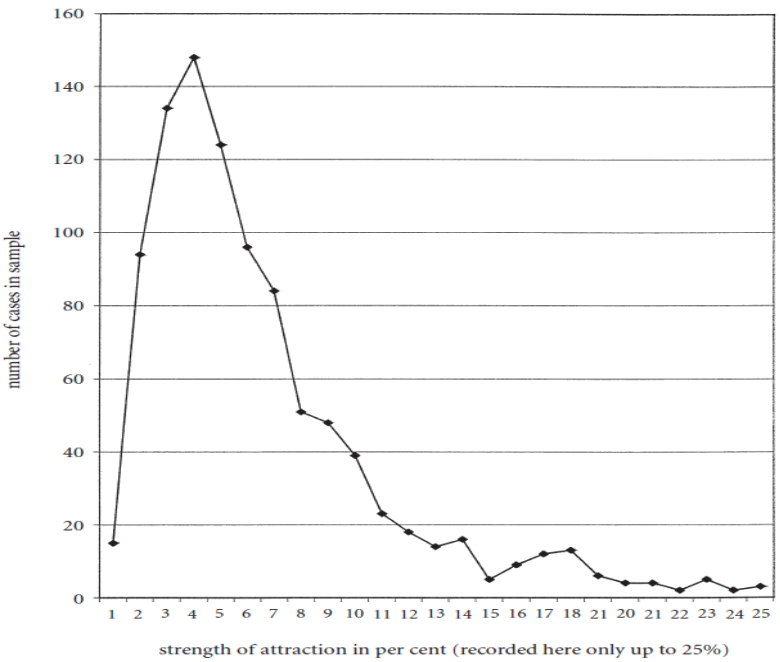

Figure 1. Strength of attraction between node and top collocate (two single word-forms)

Attraction between word-forms, lemmas and lexical sets

The findings so far concern only the collocation of pairs of word-forms. However, such a very simple (and simplistic) concept of collocation will considerably underestimate the strength of attraction between words. Consider again the example with implemented: such a calculation would mean that *implemented a plan, implemented a policy,* and *implementing a plan* would all be counted separately, although to a human analyst all three phrases are part of a single semantic pattern.

Lemmatizing the collocates list occasionally makes a considerable difference to the calculation of strength of attraction:

– mistakes <made 20%>; mistakes <MAKE 43%>

However, lemmatization often makes less of a difference than one might think. First, forms of a lemma often differ greatly in frequency, so it is rare for several different forms of a lemma all to occur amongst the top 20 collocates. Second, different forms of a lemma often have different collocational behaviour. What regularly makes a much larger difference is to group the collocates into sets of approximate synonyms:

– aerial <bombardment 5%, bombing, attacks> 12%
– obey <orders 10%, order, law(s), rules, command(s), instructions> 38%

It would be difficult to quantify this effect precisely, since it would be difficult to get a consensus between observers as to exactly which words should be grouped in this way. For example, there are many cases where the collocates obviously share a semantic feature, but are equally obviously not synonyms:

– economically <politically, socially, culturally, militarily>
– forehead <hand, eyes, nose, face, chin, cheeks, lips>

And there are cases where a word has two senses, and two corresponding sets of collocates from different semantic fields:

– commanded <respect, attention>
– commanded <army, troops, forces>

– hip <hop, rap, soul, jazz, music>
– hip <knee, hand, back, bone, shoulder, leg, thigh>

What is clear is the overall strength of lexical attraction involved. If we look only at node and top collocate, then a large majority of nodes have a strength of

attraction of between 2 and 10 per cent. If we look at the semantic preference of the nodes, defined as the relation between nodes and sets of collocates from a well-defined semantic field (including approximate synonyms), then the mode on the graph in Figure 1 would shift considerably to the right.

Topic and text-type

In order not to exaggerate the strength of attraction between words, I am presenting results at each stage of the argument in ways which tend to underestimate collocational attraction. Another source of under-estimation is built into the figures, since they are averaged across a whole corpus, and therefore take no account of different expectations in different texts and text-types. For example, *kidney* occurs in two different topical contexts, as illustrated by two sets of collocates:

– kidney <failure 9%, disease, transplant(s), patients, dialysis, problems, disorders, blood, cancer>
– kidney <steak 6%, beans, pie>

Although the collocation *kidney <failure>* is more frequent in the data-base, in a recipe we would nevertheless expect *kidney <steak>*. Conversely, in a recipe *kidney <problems>* is unlikely, but not impossible. The variation in collocations in different topics and text-types is beyond the scope of this article.

Semantic relations

The frequency figures presented above are surface evidence of semantic patterns. No set of twenty top collocates is a random list, and there are always semantic relations between node and collocates, and amongst the collocates themselves. Amongst the collocate pairs with the strongest attraction are many fixed phrases, including idioms, and many compound nouns, including (quasi-)technical terms, such as

– barbed wire, fairy tale, hay fever, managing director, saturated fats, stumbling block, waste disposal

Amongst other frequent patterns, there are many cases of co-occurring antonyms as in (a) (in all examples given, the antonym occurs amongst the top five collocates), co-hyponyms as in (b), hyponym and superordinate as in (c), terms for member and group as in (d), and approximate synonyms (less frequent) as in (e):

– (a) alive <dead>; ancient <modern>; bad <good>; black <white>; bottom <top>; cold <hot>; dark <light>; dry <wet>; late <early>; false <true>; inner <outer>; inside <outside>; left <right>; low <high>; minor <major>; passive <active>; poor <rich>; short <long>; soft <hard>
– (b) bowls <plates, vases, jugs, glass>; shirts <jeans, shorts, jackets, ties, trousers, caps, suits>
– (c) buses <transport>; cholera <disease>
– (d) aunt <family>; cattle <herd>
– (e) ashamed <embarrassed, guilty>; towns <cities>; praying <hoping?>

Using different methods, Justeson and Katz (1991) and Fellbaum (1995) have shown that antonym pairs (admittedly an ill-defined relation) co-occur much more frequently in text than would be expected by chance.

Some restrictions on collocations are purely lexical, hence *high school* and not **superior school* (compare French *école supérieure*), and the more frequent collocation of *top-bottom* rather than the less frequent *top-foot*. In other cases, the co-occurrence of words is due to the co-occurrence of things in the world (e.g. *aunt <uncle>; cakes <biscuits>; rivers <lakes>*). Content words would not frequently co-occur unless they shared some semantic feature, and they would not co-occur in individual texts unless they stood in some semantic relation to each other and were contributing to a cohesive text. So, collocation is sometimes motivated by real world facts and by semantics. A detailed analysis of these semantic relations is also beyond the scope of this article (but see Melčuk 1996 for one detailed proposal for studying such relations).

A second method: chains

The first part of this paper has defined collocation as the habitual co-occurrence of two unordered content words, or of a content word and a lexical set. This shows that co-selection of content words within a small span is the general rule, but says nothing about the actual form of the phrases in which the pairs occur. The second part of the paper describes a method of extracting from a corpus phrases which consist of a combination of grammatical and content words, and identifies some of the most frequent phrases in the language in this sense.

The frequency of words and phrases

Words are very unequal in frequency: a few words are very frequent, whereas most words are very rare. In a typical individual text or in small corpora of one million words or so, up to half the words will occur only once each. As Kilgarriff (1997: 135) puts it in an excellent review of the area, ‘a central fact about a word

is how frequent it is', since frequency relates to several other features of words and their meanings. Frequent words tend to be shorter, and to be irregular in morphology and spelling. There is a high correlation between very high frequency words (roughly the top 100) and what are traditionally known as grammatical words (though see below on this distinction). High frequency words are, by definition, more predictable (we can understand telegrams in which grammatical words have been omitted). They also have more meanings (i.e. are ambiguous out of context): this was shown by Zipf (1945), and is evident in a rough and ready way from a glance at a dictionary, where short frequent words often have many column inches devoted to them, whereas longer and rarer words tend to be more specific or specialized in meaning.

So, word frequency (facts about word tokens in *parole*) relates to several aspects of the lexicon (word types in *langue*), and word frequency lists are therefore one of the most important kinds of information derivable from corpora. However, there are only the beginnings of corresponding frequency lists of phrases. This is unfortunate, because, as Summers (1996: 262–263) puts it: 'some of the most frequent words in the language [... are] not frequent by virtue of their single word uses [...] but because they often occur in so many set phrases or chunks'. Sinclair (1999: 162) makes the same point, that frequent words play a major role in the composition of recurrent phrases. For example, there are many multi-word units which are themselves frequent, and which contain high frequency grammatical words, such as

– at least; because of; in case of; in order to; of course

Any work in this area has to reach a compromise between what is desirable and what is possible. Moon (1998: 44) uses a sophisticated definition of multi-word units, which emphasizes their abstract and variable form and their functions in expressing evaluative social connotations, contributing to text cohesion, and so on. She does provide much frequency data, but concludes that the complete set of multi-word units in English is 'uncharted, unquantified, and indeterminate'. Rather than attempting to identify multi-word units automatically in corpora, she uses existing published lists as a starting point, and then searches corpora for these known units and their variants (Moon 1998: 45). It is only by using a very much more restricted definition of multi-word unit, that Biber et al. (1999: 990–1024) can provide an automatic retrieval method for units, and start to list, for different text-types, the most frequent 3-, 4- and 5-word 'lexical bundles' (see below Section 3.3 on this term and on a retrieval method).

The example of *way*

The frequency-meaning relation is particularly relevant for phraseology, since, when they occur in phrases, frequent words are usually not ambiguous at all. A bald statement of the principle is given by Hunston and Francis (2000: 270): 'most words have no meaning in isolation, or at least are very ambiguous'. For example, *way* is one of the most frequent words in English (usually in the top 100 in lemmatized lists), and gets nearly 50 column inches in the Cobuild (1995b) dictionary, under 94 sub-headings. But it is frequent not because people make frequent references to a *way* in the sense of "path or road", but because it occurs in many phrases where it is delexicalized and has only residual relations to this historically original denotation. (This is another area in which terminology is not standardized. Different terms which often mean the same include 'delexicalization', 'desemanticization', 'semantic bleaching', 'semantic weakening', 'feature sharing' and 'co-selection'.)

Where it does denote a physical path, it usually requires support from another content word (e.g. in compounds such as *railway*, *highway*). It can be used to mean "route" (*what's the best way to get to London?*), but this can be interpreted abstractly (*I'd take the train, if I were you*). Its complex semantic and pragmatic history is recorded in the Oxford English Dictionary (OED 1989), via different senses from concrete "path", to more abstract "passage" (*doorway, way out*) and "method" (*the best way to do it*), to temporal and discourse uses. In a (very small) random sample of 200 examples of *way* from 150 million words, very few were interpretable locationally or in terms of travel. There was only one example of reference to a concrete road, and a few other instances of travel:

– trudged down [a] great new processional way
– travelling the whole way [...] from southern Ireland

Thus, *way* has a wide range of positional and temporal meanings, plus discoursal and purely idiomatic uses, but vanishingly few uses where the sense is "path or track":

– position: *the other way round*
– method: *the correct way of holding it*
– temporal: *always; all the way through the film*
– adverbial: *away*
– concessive: *in a way*
– discourse marker: *by a long way; anyway; by the way*
– not fully transparent idioms: *with her all the way; rub someone up the wrong way; in an open way; there was no way they were going to do that*

For more detailed analyses of *way*, see Sinclair (1999: 166–172), who relates the analysis to a discussion of the most frequent words in English, and Stubbs (2001: 206–209), who reviews the large literature on the almost completely delexicalized uses of *way* in the *MAKE one's way* construction. For a comparable analysis of 'stabilized expressions', which include the word *time*, see Lenk (2000).

Cases of this type are discussed in the large literature on grammaticalization, where data from many languages show regular diachronic development of place words into temporal and discourse expressions. Often the semantic content is weakened (delexicalization) and the pragmatic meaning is strengthened (Traugott & Heine 1991). As just one single further example, compare the diachronic development of German *Weg: Weg* ("way, path"); *weg* ("away"); *wegen* ("due to, because of").

These observations have several implications. (1) If such phrases are themselves frequent, then the decision to treat them as units, or not, will affect the frequency of their constituent orthographic words in word frequency lists. These lists present frequencies which are partly the result of something else: the frequency of phrases which contain the words. (2) Those highly frequent words which are often regarded as content words may be rarely used with their full lexical meaning: the boundary between content and grammatical words may be less certain than is often assumed (Sinclair 1999: 159). (There are of course morphological differences, such as the lack of inflections on grammatical words, but any semantic distinction is probably doomed to failure.) (3) Words may show a range of uses in the contemporary language which are the result of diachronic changes. These changes often lead to increasing abstraction: from physical place terms to temporal terms to discourse terms. Such delexicalization is a logical consequence of their frequent use in phrases, where meaning is dispersed across the phrase as a whole. All of these points are further illustrated below.

'Chains'

I now take a conceptually very simple definition of phrase, describe a method of extracting frequent strings of word-forms from corpora, and then discuss the lexis which occurs in these strings.

Corpus methods make it possible to observe repeated events in language use, and one type of repeated event is a recurrent 'chain' of word-forms. A 'chain' is defined here as a linear sequence of uninterrupted word-forms, either two adjacent words, or longer strings, which occur more than once in a text or corpus. There are no standard terms for such strings, which are called 'dyads', 'tryads', etc. by Piotrowski (1984: 93), 'clusters' by Scott (1997: 41), 'recurrent word-combinations' by Altenberg (1998: 101), 'statistical phrases' by Strzalkowski

(1998: xiv), 'lexical bundles' by Biber et al. (1999: 993), or simply n-grams. There are also no standard terms for the abstract grammatical sequences which underlie such strings. With different definitions and different emphasis on lexis, grammar and/or pragmatics, terms include 'canonical form', 'construction', 'extended lexical unit', 'frame', 'pattern' and 'template'. Here, I mainly use the term 'pattern', though I appreciate that this term is given a more restricted meaning by Hunston and Francis (2000: 1).

Such strings have a different status in texts and corpora. In an individual text, recurrent chains contribute to textual cohesion, and are one measure of how repetitive a text is. In a corpus, chains which occur frequently, in different texts from different speakers, may provide evidence about units of language use. Note that I have not said that chains are units in the language, since many of the chains identified below are not complete syntactic or semantic units (Piotrowski 1984: 97 calls them 'linguistic half-products'). These aspects of repetition and cohesion in individual texts, and their relation to wider intertextual patterns in the language, have hardly been studied (and only some aspects will be discussed here).

A program was written to identify frequent phrases in this sense (see Acknowledgements). The program can be given an input file of any length. As an illustration, suppose the file starts (as does Chomsky's *Aspects of the Theory of Syntax*):

> This study will touch on a variety of topics in syntactic theory and English syntax, a few in some detail, several quite superficially, and none exhaustively.

If the program is asked to identify recurrent 3-word chains, then it proceeds through the text, with a moving window, identifying and storing each 3-word string:

– this_study_will, study_will_touch, will_touch_on, touch_on_a, on_a_variety, a_variety_of, etc.

Each new string is checked against stored strings, and the program prints out, with their frequencies, those which occur more than once. In this case, it might provide (here purely hypothetical illustrative figures) a list starting:

– 25 a_variety_of
– 20 in_some_detail

Some strings will be fragments which have no obvious grammatical or semantic status. For example, one string identified in the case above would be *superficially_and_none*. However, chains which recur frequently are of more

interest. The Appendix lists the 140 5-word chains occurring 10 times or more in a corpus of 2.5 million words.

The 2.5 million words were constructed from three small reference corpora, LOB, FLOB and LUND, that is, respectively one million words of written language, published in 1961 (prepared under the direction of Geoffrey Leech and Stig Johansson), one million words of written language, published in 1991 (prepared under the direction of Christian Mair), and 500,000 words of spoken language (prepared under the direction of Randolph Quirk and Jan Svartvik). The corpus therefore contained over 1,000 text samples of British English, of between 2,000 and 5,000 running words each, from a wide variety of spoken data (about 20 per cent), casual and more formal written data, both fiction and non-fiction (about 80 per cent).

Some frequent 5-word chains

To illustrate the method, I will take examples of 5-word chains. In this corpus, the chain *at the end of the* was over twice as frequent as any other 5-word chain, and *end* occurs in three different chains in the top dozen. Out of the top ten chains, five have the structure: *PREP the NOUN of the*. Out of the top 45 chains, which all have frequencies of between 42 and 6 per million running words, 13 have this structure. They are listed in descending frequency in [1]:

– [1] at the end of the
in the middle of the
in the case of the
at the beginning of the
by the end of the
at the top of the
at the time of the
on the part of the
at the bottom of the
on the edge of the
towards the end of the
in the centre of the
on the basis of the

Further chains in the Appendix illustrate closely related patterns, for example:

– the other side of the; on the other side of; on either side of the; at the end of a; etc.

The chains program operationalizes a concept of repeated units. However, the list of recurrent chains, which is produced automatically without the intervention of the analyst, is an intermediate representation, which does not itself pick out linguistic units, but only presents data in a way which helps the analyst to identify units. Indeed, the variants of the chains are evidence of units which are more abstract than merely uninterrupted strings of unlemmatized word-forms. It is clear (to the human analyst) that the chains in [1] have very similar syntax and semantics. Not only do they fit the grammatical pattern noted above. In addition, most of the nouns are terms for place or time, and some can be used both for physical places and time periods, e.g.:

– at the end of the pier / of the morning
– in the middle of the room / of the night
– at the beginning of the chapter / of the month

Most denote the outer limit or the middle of areas of space or periods of time (e.g. *end, edge, middle, centre*). Some (see the longer list in the Appendix) can be either body parts or place terms (*back, bottom, foot, side*). They all seem to have intuitively clear core meanings out of context, but, as with all frequent words, they have a range of uses which are perhaps less intuitively obvious. For example, *end* is given 24 column inches and 40 different sub-sections in the Cobuild Dictionary (1995b).

We find exactly what Hunston and Francis (2000: 96) find much more generally: that a pattern occurs with a restricted lexical set; a few words in the set occur very frequently; other semantically related words occur more rarely. Native speakers can make intuitive judgements as to which words would be acceptable in the frame, and this allows variability and creativity.

Although the 2.5 million word corpus consisted of several hundred diverse text samples, it is important to check that the patterns are not due to some idiosyncrasy of this corpus. One indication that the patterns are not an artefact of the small initial corpus is that Biber et al. (1999: 1015) identify *at the end of the* as the most frequent 5-word 'lexical bundle' in academic prose, and identify as characteristic of academic prose, the 4-word grammatical frame *the NOUN of the*, where frequent nouns include

– end, beginning, top, edge, centre, part (all in [1])
– base, position, shape, size, start, structure, surface, form, length, magnitude, composition, temperature, level, context, rest

Chains of the pattern *PREP the N of the* are frequent in written data. However, *at the end of the* is also the most frequent 5-word chain in the LUND corpus (500,000 words of spoken English), along with other chains with the same pattern,

and the pattern is not restricted to academic prose, as Biber et al. (1999) seem to imply. Also in the top 100 in LUND are: *at the beginning of the, on the far side of, at the bottom of the, at the foot of the*.

I also compared the frequencies of the chains in [1] with their frequencies in the 100-million word British National Corpus (BNC) (90 million words of written and 10 million words of spoken English). Normalized to estimated occurrences per million words, the frequencies in the two corpora, the small 2.5 million word corpus and the 100 million word BNC respectively, are as follows:

at the end of the	42	45
in the middle of the	19	16
in the case of the	15	9
by the end of the	13	18
at the beginning of the	13	9
at the top of the	12	11
at the time of the	11	12
on the part of the	11	8
at the bottom of the	10	7
on the edge of the	8	7
towards the end of the	8	8
in the centre of the	6	6
on the basis of the	6	6

The frequencies are remarkably similar. So, we can be confident that these 5-word chains are not an artefact of the small corpus which I started with, but are frequent 5-word chains in general English, though some are more frequent in written genres.

In summary so far: We have a method of identifying recurrent uninterrupted strings of unlemmatized word-forms. The chains listed were identified purely on grounds of raw frequency, though the grammatical and semantic patterns were not identified automatically. The method will find a chain such as *on the top of the*, but will not count this together with variants such as *on the very top of the* or *on top of the*. This is an obvious limitation, but 'chains' are only one kind of evidence in a quantitative study of phraseological units, and provide only one method of identifying repeated events across corpora and one definition of phrase.

Other patterns

The Appendix also contains examples of other patterns, including several chains with discourse functions:

– as a matter of fact; as a result of the; from the point of view; it seems to me that; as far as I know; there can be no doubt; but on the other hand; it is clear that the

Altenberg (1998: 117) gives further examples. Other chains, though a minority, signal the topic and/or text-type of individual text samples in the corpus. For example, given two sets of chains such as the following, there is no doubt which come from academic articles and which from novels:

– in the context of the; it can be seen that; in the case of the; it is interesting to note
– to have a word with; there was no reason why; I've never done this before; in the back seat of

Similarly, these chains, even in isolation, allow rather accurate guesses as to their source:

– ask the minister of agriculture; the right hon and learned; the book of common prayer; of violence against the person

Indeed some individual texts are much more repetitive than others, and many readers will immediately recognize the source of these 5-word chains in the Bible:

– and it came to pass; the word of the Lord; verily I say unto you; the angel of the Lord

The formulaic nature of certain text-types is discussed by Youmans (1990), who uses different measures of repetition in texts.

Explanations?

Could it be that the frequency of chains such as *at the end of the* is just an automatic consequence of their high frequency constituent words? The words *the* and *of* are the two most frequent words in the language; *in*, *on* and *at* are usually in the top 20 words; and *end* is one of the few content words which occur in the top 200 words in frequency lists. Since these words are themselves highly frequent, they have a good chance of co-occurring, and the frequency of such chains might be partly a consequence of this plus the analytic syntax of English. (For example, in corpora of French and German, the pattern would be less clear because of the gender system and the resultant allomorphic variation in definite articles and in article plus preposition: e.g. German *der, die, das, in dem, im*, etc.)

However, as noted above, it is not that words are frequent and therefore tend to co-occur. It is precisely because they are part of frequent phrases that they co-occur frequently. Grammatical words do not occur on their own: their function is to form larger units.

In addition, a probability calculation does not explain the patterns. In the version of the LUND corpus which I have used, the number of running word tokens is 498,183. The frequency of *the* is 21,171, of *of* 11,307, and of *of the* 1,954. At a random point in the corpus, the probability of *the* being the next word is 21,171/498,183 = 0.04. Similarly, the probability of *of* is 0.02. The expected probability of *of* and *the* occurring next to each other is therefore 0.04 × 0.02 = 0.0008. And the probability of them occurring in the sequence *of the* is half of that: = 0.0004. But the observed probability is 1,954/498,183 = 0.004. That is, the observed frequency of *of the* is ten times higher than would be expected by chance. In addition, 8 per cent of occurrences of *end* are in the chain *at the end of the*. And 24 per cent of the occurrences of *at the NOUN of the* have *end* as the noun. Finally, the list of the top 200 words in a large corpus (Sinclair 1999: 176–177) does contain several time and place words, including *end*, but otherwise does not correspond at all to the words identified in the chains above. The top content words, in descending frequency, are:

– said, new, time, people, year, first, last, years, back, think, way, right, world, say, work, life, own, long, man, week, come, yesterday, next, little, want, today, women, same, end, place

Ultimately, the frequency of sequences such as *PREP the end of the NOUN* is explained by the fact that it fits into a preferred phrasal schema in English, and by the fact that this is the kind of thing that speakers frequently talk about. This requires a social explanation.

On lexis, grammar and semantics

As part of their argument that 'the normal use of language is to select more than one word at a time', Renouf and Sinclair (1991) recommend studying collocational frameworks, which they define as discontinuous sequences of two grammatical words, 'somewhere between a word and a group' (p. 129), for example *a ? of* (as in *a lot of ; a couple of ; a pint of*) or *too ? to* (as in *too late to; too much to; too young to*). They document the 'tendency of these frameworks to enclose characteristic groupings of words' (p. 128). In the frame *at the ? of the*, the top nouns include

– end, beginning, bottom, centre, top, foot, time, back, side

Of course, if we search for this frame, then we just find again the nouns in [1], but we also find other frequent nouns which mean either "outer edge of" or "centre of", such as

– conclusion, door, entrance, fringe, head, mouth, onset, outbreak, start, threshold, heart, navel.

Similarly, in the frame *on the ? of the*, the top word-forms and other less frequent word-forms which mean roughly "edge" are

– edge, side, top, morning, bottom, eve, back
– circumference, fringe, front, lip, outskirts, periphery, roof, surface, tip

Some words are partly delexicalized in such uses. For example, *eve* (in *on the eve of the war*) means "immediately before", and has lost its literal meaning of "evening". Compare *turn* (*at the turn of the century*), and *lip* (*on the lip of the crater*). In such cases, the unit of meaning is the whole phrase: the meaning cannot be inferred with complete accuracy (that is, compositionally) from the constituent words.

A technique to identify the characteristic syntactic frames in which words occur is to identify their most significant collocates as measured by a t-score (i.e. frequency of co-occurrence, corrected for the frequency of the individual constituents: Stubbs 1995: 36–38). In the 50-million word CobuildDirect corpus, all of the following have *the* and *of* and either *at*, *in* or *on* as their top three grammatical collocates (often as the top three collocates). That is, they most frequently occur in the frame *at/in/on the ? of the*:

– beginning, bottom, centre, edge, end, middle, side, top

This allows us to establish the core lexis for the grammatical frame.

Conclusions

The long term aim of the work presented here is to describe 'all the frequently-occurring items in the language in a principled way' (Hunston & Francis 2000: 14). These items can be described in terms of the co-selection of lexis and grammar. The paper has proposed two methods of inspecting a corpus in a single process, as Krishnamurthy (2000: 41) puts it, in order to state phraseological generalizations which are valid for the whole lexicon. First, I proposed a method of studying pairs of collocates which frequently co-occur because they share a meaning. (If they did not share a meaning, then the texts in which they occur would not be coherent.) Second, I proposed a method of studying frequent

grammatical frames and the lexis which occurs in them. This reveals sets of semantically related lexis, and reveals patterns which are integrated form-meaning pairs.

The two methods capture different aspects of the phraseology of English. (1) The first method uses the concept of collocation (defined as habitual co-occurrence), and identifies frequent co-selections of two content words within a small span. In this case, the (unordered) pairs of words can be discovered automatically, and generalizations can be drawn about the frequency and strength of attraction within the pairs. However, to discover the semantic relations between the words, there is no alternative to examining each pair individually. (2) The second method involves colligation (the relation between content and function words, and between words and grammatical categories), and identifies frequent co-selections of a content word and an associated grammatical frame. In this case, the most frequent 'chains' can be discovered automatically, but generalizations about the constituent lexis still require manual analysis.

The present paper is largely methodological, and illustrates how systematic observation of large data sets can allow generalizations about phraseology. Work in progress will present analyses of three topics which have been mentioned here only in passing: the semantic and pragmatic features of frequent collocations and multi-word chains; the variation of collocations and chains across different text-types; and the functions of extended lexical units in textual cohesion.

Acknowledgements

I am most grateful to my student research assistants in Trier, past and present, for help with data preparation and analysis: Isabel Barth wrote the 'Chains' program and provided ideas and references; and Oliver Mason, Kerstin Günther, Christine Spies and Bettina Starcke extracted data from corpora and data-bases. For comments on earlier versions of the paper, I am grateful to Andrea Gerbig and Mike Scott, and to seminar audiences at UMIST (University of Manchester Institute of Science of Technology), UK, in March 2001, and at the University of Basel, Switzerland, in May 2001.

References

Altenberg, B. (1998). On the phraseology of spoken English: the evidence of recurrent word-combinations. In A. P. Cowie (Ed.), *Phraseology* (pp. 101–122). Oxford: Oxford University Press.

Biber, D., S. Johansson, G. Leech, S. Conrad, & E. Finegan (1999). *Longman Grammar of Spoken and Written English*. London: Longman.

Bolinger, D. (1976). Meaning and memory. *Forum Linguisticum*, 1 (1), 1–14.
Cobuild (1995a). *Collins COBUILD English Collocations on CD-ROM*. London: HarperCollins.
Cobuild (1995b). *Collins COBUILD English Dictionary*. London: HarperCollins.
De Beaugrande, R. (1999). Reconnecting real language with real texts: text linguistics and corpus linguistics. *International Journal of corpus Linguistics*, 42 (2), 243–259.
Fellbaum, C. (1995). Co-occurrence and antonymy. *International Journal of Lexicography*, 8 (4), 281–303.
Fox, G. (1987). The case for examples. In J. Sinclair (Ed.), *Looking Up* (pp. 137–149). London: Collins.
Hunston, S., & G. Francis (2000). *Pattern Grammar*. Amsterdam: Benjamins.
Hymes, D. (1962). The ethnography of speaking. In J. A. Fishman (Ed.), *Readings in the Sociology of Language* (pp. 99–138). The Hague: Mouton.
Justeson, J. S., & S. M. Katz (1991). Redefining antonymy: the textual structure of a semantic relation. *Using Corpora.* Proceedings of 7th Annual Conference of the UW Centre for the New OED and Text Research (pp. 138–153). Oxford.
Kilgarriff, A. (1997). Putting frequencies in the dictionary. *International Journal of Lexicography*, 10 (2), 135–155.
Krishnamurthy, R. (2000). Collocation: from *silly ass* to lexical sets. In C. Heffer & H. Sauntson (Eds.), *Words in Context* [CD-ROM], Discourse Analysis Monograph 18. Birmingham: University of Birmingham.
Lenk, U. (2000). Stabilized expressions in spoken discourse. In C. Mair & M. Hundt (Eds.), *Corpus Linguistics and Linguistic Theory* (pp. 187–200). Amsterdam: Rodopi.
Melˇcuk, I. (1996). Lexical functions: a tool for the description of lexical relations in a lexicon. In L. Wanner (Ed.), *Lexical Functions in Lexicography and Natural Language Processing* (pp. 37-102). Amsterdam: Benjamins.
Miller, G. (1998). Foreword. In C. Fellbaum (Ed.), *WordNet: An Electronic Lexical Database* (pp. xv–xxii). Cambridge, MA: MIT Press.
Moon, R. (1998). *Fixed Expressions and Idioms in English. A Corpus-Based Approach*. Oxford: Clarendon.
OED (1989). *The Oxford English Dictionary*, 2nd edition, 20 vols., ed. by J. A. Simpson & E. S. C. Weiner. Oxford: Clarendon.
Pawley, A., & F. H. Syder (1983). Two puzzles for linguistic theory. In J. C. Richards & R. W. Schmidt (Eds.), *Language and Communication* (pp. 191–226). London: Longman.
Piotrowski, R. G. (1984). *Text, Computer, Mensch*. Bochum: Brockmeyer.
Renouf, A., & J. Sinclair (1991). Collocational frameworks in English. In K. Aijmer & B. Altenberg (Eds.), *English Corpus Linguistics* (pp. 128–144). London: Longman.
Scott, M. (1997). *WordSmith Tools Manual*. Oxford: Oxford University Press.

Sinclair, J. (1991). Shared knowledge. In J. Atlatis (Ed.), *Linguistics and Language Pedagogy* (pp. 489–500). Georgetown: Georgetown University Press.
Sinclair, J. (1999). A way with common words. In H. Hasselgard & S. Oksefjell (Eds.), *Out of Corpora* (pp. 157–179). Amsterdam: Rodopi.
Strzalkowski, T. (Ed.). (1998). *Natural Language Information Retrieval*. Dordrecht: Kluwer.
Stubbs, M. (1995). Collocations and semantic profiles: on the cause of the trouble with quantitative studies. *Functions of Language*, 2 (1), 23–55.
Stubbs, M. (2001). *Words and Phrases: Corpus Studies of Lexical Semantics.* Oxford: Blackwell.
Summers, D. (1996). Computer lexicography: the importance of representativeness in relation to frequency. In J. Thomas & M. Short (Eds.), *Using Corpora for Language Research* (pp. 260–266). London: Longman.
Traugott, E., & B. Heine (Eds.). (1991). *Approaches to Grammaticalization*. 2 volumes. Amsterdam: Benjamins.
Wray, A. (2002). *Formulaic Language and the Lexicon*. Stanford, CA: Cambridge University Press.
Youmans, G. (1990). Measuring lexical style and competence: the type-token vocabulary curve. *Style*, 24 (4), 584–599.
Zipf, G. K. (1945). The meaning-frequency relationship of words. *Journal of General Psychology*, 33, 251–256.

Appendix

Five-word chains occurring 10 times or more in a 2.5 million word corpus.

>> marks chains with the pattern: PREP the NOUN of the
> marks chains with related patterns.

>> 104 at the end of the
>> 48 in the middle of the
> 40 the other side of the
>> 37 in the case of the
36 and at the same time
33 as a matter of fact
33 as a result of the
>> 33 at the beginning of the
>> 33 by the end of the
33 for the first time in
>> 29 at the top of the
>> 28 at the time of the
>> 27 on the part of the

>> 25 at the bottom of the
25 in the house of commons
> 25 the turn of the century
24 from the point of view
24 the point of view of
> 23 on the other side of
22 in the same way as
22 it seems to me that
22 of agriculture fisheries and food
22 there is no doubt that
20 all the rest of it
> 20 in the form of a
20 on the other hand the
19 and all the rest of
19 as far as I can
>> 19 on the edge of the
>> 19 towards the end of the
18 at the same time the
18 is one of the most
18 no no no no no
18 this is one of the
> 17 at the end of a
17 in such a way as
> 17 the second half of the
16 for the first time the
16 go on to the next
>> 16 in the centre of the
16 it may well be that
>> 16 on the basis of the
16 thank you very much indeed
> 16 the end of the year
16 the secretary of state for
15 ask the minister of agriculture
15 for the first time since
15 it is not surprising that
15 minister of agriculture fisheries and
> 15 on the far side of
15 such a way as to
> 15 the far side of the
15 the minister of agriculture fisheries
15 the right hon and learned
15 to ask the minister of
14 as far as I know
>> 14 at the back of the

	14 but on the other hand
>	14 for the rest of the
>>	14 in the direction of the
	14 in the house of lords
>>	14 in the light of the
	14 it is clear that the
	14 on the one hand and
	14 on the other hand it
	14 on to the next question
	14 the book of common prayer
>	14 the end of the war
>	14 the first half of the
	14 to be found in the
	14 what are you going to
	13 as in the case of
>>	13 at the foot of the
>	13 in the case of a
	13 in the first world war
	13 of the house of commons
	13 of violence against the person
	12 and so on and so
	12 are you going to do
>>	12 at the expense of the
>>	12 at the turn of the
	12 crimes of violence against the
>	12 far side of the field
	12 in the course of the
>	12 in the second half of
	12 it was the first time
	12 it would have to be
>	12 on either side of the
	12 pip pip pip pip pip
	12 right hon and learned gentleman
	12 the way in which the
	12 there can be no doubt
	12 to the next question from
	11 he was one of the
	11 I mean I don’t know
>	11 in the context of a
>	11 in the early years of
	11 increase in the number of
	11 is not to say that
	11 it is not possible to
	11 m m m m m

	11	nothing to do with the
	11	of r joshua b levi
	11	of the church of england
>>	11	on the back of the
	11	on the other hand there
>>	11	on the side of the
	11	one of the things that
>	11	the far end of the
>	11	the other end of the
>	11	the rest of the world
	11	there was no sign of
>>	11	to the end of the
	11	while at the same time
	10	an hour and a half
>	10	and the rest of the
	10	as a result of a
>	10	at the other end of
	10	but at the same time
	10	do you want me to
	10	due to the fact that
	10	for a long time and
	10	I would have thought that
	10	if he will make a
	10	in and out of the
	10	in the church of england
>>	10	in the context of the
	10	is to be found in
	10	it is not easy to
	10	it should be remembered that
	10	it was going to be
	10	it was one of the
	10	it would have been a
	10	no I don’t think so
	10	the effects of noise on
>	10	the end of the first
	10	the extent to which the
	10	the fact that he was
	10	to the secretary of state
	10	was one of the most
	10	we go on to the