Functions of Language (2003) 10, 1: 61-104. [Pre-publication version 1 May 2003.]

USING RECURRENT PHRASES AS TEXT-TYPE DISCRIMINATORS: A QUANTITATIVE METHOD AND SOME FINDINGS

Michael Stubbs and Isabel Barth
University of Trier, Germany

ABSTRACT

This paper describes a method of comparing routine language use in different corpora, and presents results from a quantitative study of the norms in different text-types. To illustrate the method, we use three sub-corpora, of over 600,000 running words each, from fiction, belles lettres and academic writing. First, we confirm well-known findings that individual word-forms are differently distributed in these three text-types. Second, and more originally, we show that chains of word-forms (recurrent multi-word units or n-grams) can also accurately discriminate between the text-types, as regards both the actual chains which occur, and also the overall repetitiveness of the text-types. Third, we discuss how these methods provide evidence of previously unobserved patterns of language use and of linguistic units which are often ignored in lexical and syntactic studies.

1. INTRODUCTION

One of the many dualisms which has characterized linguistics over the past hundred years is the tension between creativity and routine in language, and linguists have struggled to find ways of appropriately stating the balance between these two opposing forces. In terms of everyday language use, there is a tension between the need to express new information and the need to maintain redundancy: texts which are too creative are difficult to understand, but texts which are too repetitive may be full of boring clichés. In terms of theory, scholars have emphasized one or other of these aspects of language. Chomskyan linguistics has always emphasised creativity, and Descartes' problem (as Chomsky 2000 calls it) is seen as central for linguistic theory. However, Firth, Halliday and Sinclair emphasize the importance of routine language. In famous statements, Firth (1935: 71) talks of the habitual, customary and typical usage of words and phrases. Halliday (1978: 4) points out, in a similar vein, that "a great deal of discourse is more or less routinized", and that we "express the same opinions over and over again". And Sinclair (1991: 109) proposes "two different principles of interpretation", the idiom principle and the open choice principle, and argues that semi-preconstructed phrases are "far more pervasive" (p.111) than is often realised. The approach to corpus study which derives from this neo-Firthian tradition (especially via Sinclair 1991) has conclusively shown the pervasiveness of recurring phraseology, as have key publications in other traditions (e.g. Pawley & Syder 1983 and Fillmore et al 1988) and overviews (e.g. Wray 2002).

In this article we show that different text-types are repetitive in different ways and to different extents. After initial discussion of word frequency data, we concentrate on multi-word strings of two or more uninterrupted word-forms 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 (1997a), "recurrent word-combinations" by Altenberg (1998), "statistical phrases" by Strzalkowski (1998: xiv), "lexical bundles" by Biber et al (1999), or simply n-grams. We call them "chains". We show that individual words and chains distinguish different text-types, and we identify some frequent word-chains in general English. Our main methods require only raw untagged text (stored as running orthographic word-forms), and rely solely on the frequency and distribution of words and chains of various lengths.

Recurrent word-chains have a different status in texts and corpora. In an individual text, they can be studied (a) as one measure of how repetitive the individual text is, and (b) as a contribution to textual cohesion (this second aspect is not discussed here). In a corpus of texts from different speakers and writers, they can be studied (a) as a predictable characteristic of different text-types, and (b) as evidence of units of routine language use. The chains themselves are not necessarily linguistic units, but only evidence of such units, since many chains are not structurally complete, and since the units which have psycholinguistic status are presumably more abstract schemas (see 9.2 and Stubbs 2001). We discuss mainly the relation between chains and intertextual patterns in the language, since this is a problem which has been studied to only a very limited extent.

2. THE DATA

Our main data are drawn from the Brown, LOB, Frown and FLOB corpora: that is, comparable one-million-word corpora of American and British English from 1961 and 1991[1]. From these standard reference corpora, we constructed three sub-corpora, which each consist of a large number of short samples and make it unlikely that stylistic idiosyncrasies of any single author or text could have biased the results. Each sub-corpus consisted of over 300 2,000-word samples, as follows:

> FICTION. The 2,000-word samples in categories K, L, M and N ("Imaginative Prose": general, mystery, detective, science, adventure and western fiction) were extracted and collated as FICTION: in total about 706,600 words[2].
>
> BELLES. The 2,000-word samples in category G ("Belles Lettres, Biography, Memoirs, etc") were extracted and collated as BELLES: in total about 610,800 words.

> LEARNED. The 2,000-word samples in category J ("Learned": natural and social sciences, humanities, etc) were extracted and collated as LEARNED: in total about 643,500 words. LEARNED clearly covers a very wide topic range, but there is empirical evidence that academic texts share a considerable sub-technical vocabulary, irrespective of subject area (Yang 1986, Flowerdew 1993, Coxhead 2000, Nation 2001: see further below).

The Brown corpus manual admits that these categories are "subjective". More accurately - and more fairly to the corpus designers - the classification relies on external functional and sociolinguistic criteria, and assumes that texts sampled in these categories are internally relatively homogeneous, but that there are significant linguistic differences between the categories. The Brown corpus was carefully designed to allow such comparisons, and even the relatively small samples from these corpora have been shown to reveal clear text-type differences (Biber 1988), though see 9.1.

We therefore have three sub-corpora: one fiction (FICTION), and two non-fiction (BELLES and LEARNED). In FICTION and BELLES both style and content are likely to be important, and in LEARNED style is likely to be subordinate to content. This second point is supported by empirical data presented below and also by different corpus methods used by Youmans (1990).

The findings presented below are limited to the 50 most frequent words and chains of different lengths in these three sub-corpora. To ensure that data from different sized sub-corpora are comparable, we normalize frequencies (rounded to the nearest 5) to estimated occurrences in one million running words. Statistical tests, where we use them, were of course carried out on the raw frequency data. In several cases, the differences between sub-corpora are so large that no statistical tests are necessary, and indeed it is sometimes more useful to carry out a test to show those cases where a difference between sub-corpora is <u>not</u> statistically significant. In sections 3 to 8 we discuss, with relatively little interpretation, the quantitative lexical findings which differentiate the three sub-corpora, and in section 9 we discuss the nature of the analytic units which we call "chains" and whether they are evidence of linguistic units.

3. THE DISTRIBUTION OF SINGLE WORD-FORMS IN DIFFERENT TEXT-TYPES

Since work by Zipf and others in the 1930s, many quantitative aspects of single words have been thoroughly discussed, including their frequency and distribution in different text-types. In this example (from Johansson 1981, cited by Kennedy 1998: 106), there is no doubt which words come from academic texts and which from fiction:

> *constants, measured, thermal, theoretically*
> *damned, impatiently, kissed, sofa*

3.1. THE TOP WORD-FORMS IN THE THREE SUB-CORPORA

The most frequent word-forms in unlemmatized frequency lists already show differences between text-types, most obviously in the occurrence, rank-order and frequency of the pronouns. In the top 50 words from the three sub-corpora, there occur, in descending frequency

FICTION: he, I, it (in top 10); his, you, she, her, him, they, my, me, we

BELLES: he (in top 10); it, his, I, they, her, their, we, she, him

LEARNED: (no pronouns in top 10); it, we, their, his, he, they

In addition, frequencies differentiate the text-types very clearly. For example, the frequencies for *he* and *she*, here estimated per million running words, are so different that no statistical test is required to show that the difference is significant.

FICTION	he 17,840	she 8,460
BELLES	he 8,830	she 2,600
LEARNED	he 2,320	she 350

The top 50 words in text collections are almost always grammatical (function) words. The only clear exception in our sub-corpora is *said* in FICTION. BELLES has *all* and *more*, and LEARNED has *other* and *such*, which are on the borderline between grammatical and lexical (content) words.

3.2. THE TOP WORD-FORMS MINUS A 100-WORD STOP LIST (100 WORD-FORMS)

Broader lexical differences become sharper if stop-lists are used to ignore words which are frequent in many text-types and are therefore not good discriminators. For example, if we use a stop-list of the 100 most frequent words in the combined corpora, then this reveals differences in content words. Using the stop-list, the content words in the top 50 are:

FICTION: know, get, go, looked, got, here, thought, right, never, again, think, come, still, little, eyes, away, came, old, went, long, going, door, face, head, room, left, something, new, hand, around

BELLES: life, world, own, work, men, both, war, great, little, still, never, long, another, art, day, later, part, same, century, American, old, family, without, fact, know, young, during, came, house, social

> LEARNED: each, used, formula, both, system, same, number, case, form, high, different, use, work, less, three, given, data, example, part, point, large, found, state, social, general, important, possible

These lists now reflect different semantic fields. FICTION has verbs of physical movement and mental activity, and nouns for parts of the body. BELLES has vocabulary for historical and cultural topics. And LEARNED has sub-technical vocabulary, that is, words which are in general usage, but more frequent in academic discourse, though not restricted to any one subject area, and often with different and specialized meanings in different subject areas. There is only a small amount of overlap between these lists and none at all among the top dozen words.

3.3. THE TOP WORD-FORMS MINUS A 2,000-WORD STOP-LIST (7,800 WORD-FORMS)

We can also use a much larger stop-list of the 2,000 most frequent "word families" (as defined by Bauer and Nation 1993): that is, lemmas with related inflected and derived forms, and therefore around 7,800 word-forms. The list used here is West's (1953) General Service List, in a version described by Nation (2001). Further clear differences emerge. FICTION has a large number of informal reduced forms, BELLES has a large number of geographical terms, and LEARNED has a large number of sub-technical words. Words in the top 30 from the sub-corpora are as follows. For FICTION it is useful to divide the list into informal reduced forms and other words:

> FICTION: don't, didn't, I'm, it's, that's, you're, I'll, wasn't, he'd, can't; job, hell, stared, nodded, finally, couple, ah, glanced, kid, frown

For BELLES it is also useful to divide the list into the place names which predominate and other words:

> BELLES: American, London, British, England, America, French, York, German, European, Europe; community, period, military, evidence, major, finally, tradition, novel, economic, policy

(These lists for FICTION and BELLES exclude several person names in the top 30.) The top words in LEARNED are mainly from the sub-technical vocabulary:

> LEARNED: formula, data, theory, evidence, individual, function, area, period, analysis, process, similar, section, areas, required, range, major, economic, available, obtained, method

Similar lists can be produced by calculating the "keywords", as defined by WordSmith Tools (Scott 1997a)[3]

3.4. GENERAL AND ACADEMIC VOCABULARY

Only very rarely can text-types be reliably identified solely by the occurrence of individual words or phrases. Abbreviated forms might reliably identify types of newspaper advertising: *2 bdrm maisonnette* is very probably from a property advertisement, and *wltm* ("would like to meet") and *gsoh* ("good sense of humour") are almost certainly from a lonely hearts ad. But "normal" words are unlikely to be reliable discriminators. Technical words may signal the topic reliably, but give only a hint about text-type. For example, the word *furlong* is certainly frequent in horse racing commentaries, but could be used in other text-types where the topic is horse racing. Individual phrases (word-chains) might be more helpful: *warm front* might be from a weather forecast, but could also be from a meteorology textbook. The co-occurrence of phrases, such as *warm front* and *in the next few days*, would be a more certain signal of a weather forecast. Conversely, there is no guarantee that these individual words or phrases will occur in a given text.

However, if we have a substantial list of words which are known to be frequent in a text-type, irrespective of subject, then their known distribution can be used as a discriminator (even if, individually, the words are not restricted to that text-type). Coxhead (2000) has produced just such a list of academic vocabulary, which consists of 570 "word families" (over 3,000 word-forms). Nation (2001) provides further discussion, and describes software which allows texts and corpora to be compared with the list. The software gives the percentage of text covered by words in the list, plus alphabetical and frequency lists of words in the input texts which are and are not in the list[4]. Coxhead's academic word-list differentiates the sub-corpora clearly in terms of percentage text-coverage:

> FICTION 1.5%; BELLES 4.9%; LEARNED 8.8%

Such figures hold out the promise of being very stable. For an independent corpus, Nation (2001) reports a very similar figure of 8.5 per cent coverage of running words for academic texts. Coxhead's list does not contain any words which are amongst the 2,000 most frequent in the language (as listed in West 1953), and conversely, the text-coverage for these 2,000 most frequent lemmas is

> FICTION 86.1%; BELLES 80.7%; LEARNED 77.1%.

We now introduce a presentational convention which makes any overlaps in lists from the sub-corpora more easily visible. Bracketed items occur in more than one list, and are therefore not, on their own, good discriminators. With rare and interesting exceptions (see 9.2), items do not occur in all three sub-corpora. Unbracketed items occur only in one sub-corpus (in the top 50 words and chains that we discuss). Thus, all three sub-corpora do contain words from the academic word-list, but different selections. The following are the top ten:

> FICTION: job, (finally), couple, (major), (area), aware, obviously, apparently, final, odd

> BELLES: community, (period), military, (evidence), (major), (finally), tradition, economic, policy, (individual)
>
> LEARNED: formula, data, theory, (evidence), (individual), function, (area), (period), analysis, process

These lists provide some evidence of grammatical differences. The FICTION list contains some adverbs, and the LEARNED list consists almost exclusively of abstract nouns.

3.5. WORDS, CHAINS, TEXTS AND TEXT-TYPES

These features of vocabulary differentiate the three text-types, but since the findings are averaged across sub-corpora, they do not necessarily characterize individual texts. For example, the pronoun *I* is less likely to occur in the LEARNED texts (and is entirely absent in some), but it occurs in the lists from LEARNED precisely because it does occur - and may be quite frequent - in other texts in this sub-corpus. Furthermore, beyond the top 2,000 or so words from general frequency lists, content words are very topic-dependent, and may therefore be frequent in an individual text, but absent in another text from the same text-type. For example, whereas no content words (apart from *said*) appear in the top 50 words from the FICTION sub-corpus, the following occur in the top 50 from a single sample of Romantic fiction narrative (FLOB P05): *duke*, *theatre*, *house*, *door*, *time*, *moor*, *park*. However, the phraseology around *I* provides a much better discriminator of text-types, and there is no doubt which of the following four-word chains come from FICTION and which from LEARNED:

> (a) I don't want to; I want you to; I don't know what
> (b) I have already mentioned; I shall show that; I will conclude with

In addition, rank orders and frequencies are very different. Set (a) come from the top 25 chains in FICTION, and all occur almost 30 times per million words. The top 50 four-word chains from LEARNED contain no personal pronouns at all. The examples in (b) are from very much further down the list, and all occur six times or fewer per million words.

So, several points must be emphasized. (1) Since content words are topic-dependent, they usually have very different frequencies in different texts. If a text is about whelks, then the word *whelks* will probably be frequent, but is likely to be entirely absent from the next hundred or thousand texts we look at. (This example is from Kilgarriff 1997, who discusses other aspects of the "whelks problem".) (2) An exception is the set of sub-technical vocabulary, which occurs in predictable percentages, not only across sub-corpora of academic writing, but also in individual academic texts. (3) It follows that content words appear higher in frequency lists from individual texts than in lists from text collections. (4) We cannot provide formal measures of the possible variability across the 300 texts in each sub-corpus, but we provide below (9.1) an informal check that our sub-corpora are internally relatively homogeneous.

4. "CHAINS"

So far we have added some findings to the large literature on the distribution of single word-forms in different text-types (e.g. Flowerdew 1993 on academic genres, Scott 1997b on newspaper articles, and Nation 2001 on teaching materials), and also demonstrated clear internal linguistic differences between sub-corpora which were established on external functional criteria. We can therefore confidently use FICTION, BELLES and LEARNED as samples to test how accurately the frequency and distribution of multi-word chains can discriminate between text-types. This has previously been investigated in only a few studies (e.g. Milton 1998, Aarts & Granger 1998, Biber et al 1999: 993-96).

4.1. THE "CHAINS" PROGRAM

Most of the data below are produced by a program which can identify "chains" in running text. The program can be given an input file of any length. Suppose the file starts, as does Bloomfield's *Language* (1933):

> "Language plays a great part in our life. Perhaps because of its familiarity, we rarely observe it, taking it rather for granted, as we do breathing or walking."

If the program is asked to identify recurrent three-word chains, then it proceeds through the text, with a moving window, ignoring punctuation, and identifies and stores each three-word string:

> language_plays_a, plays_a_great, a_great_part, great_part_in, part_in_our, in_our_life, 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:

> 50 a_great_part
> 25 in_our_life
> 20 because_of_its

As a technique for identifying linguistic units, this has clear limitations. Some strings are fragments with no grammatical or semantic status: for example, one string identified above would be *it taking it*. However, many such fragments occur only once in a large corpus, and are automatically excluded by a procedure which identifies recurrent strings. More seriously, some sequences which would probably be identified by a human analyst would be missed, since they are variants of more abstract units which contain intervening words: for example, *play a part* and *take for granted* would not be identified in the Bloomfield example, since they occur in longer chains, *plays a great part* and *taking it rather for granted*. So, there is no claim that chains are linguistic units, only that recurrent chains (a) can discriminate between text-types, and

(b) provide evidence which helps the analyst to identify linguistic units. The program operationalizes a concept of repeated units, but the list of recurrent chains is only an intermediate representation. Surprisingly many chains are idiomatic phrasal units, but variants of chains are evidence of units which are more abstract than merely uninterrupted strings of unlemmatized word-forms (see 9.2)[5].

4.2. TWO-WORD CHAINS

The top 10 two-word chains in all three sub-corpora are dominated by PREP-DET combinations such as *of the, in the* and *to the* which are the top three chains in all three sub-corpora. Even these chains distinguish text-types if their frequency is taken into account (frequencies estimated per million running words):

FICTION	of the 5390,	in the 4610	to the 3070
BELLES	of the 9895	in the 6370	to the 3300
LEARNED	of the 11,900	in the 7075	to the 3590

A χ^2-test on the original raw frequencies confirms clear differences for the overall table ($\chi^2 = 380.64$, df = 4, $p < 0.001$), and also for some of the individual comparisons, e.g. *of the* and *in the* compared for FICTION and BELLES ($\chi^2 = 81.18$, df = 1, $p < 0.001$), and for BELLES and LEARNED ($\chi^2 = 8.07$, df = 1, $p < 0.01$).

In the top 50 chains, the different distribution of personal pronouns becomes much sharper.

> FICTION: (he was 2050); (he had 1865); she was; he said; (I was 950); (in his 905); she had; (that he 910); and he; and I; they were; as he; I don't; of her; I had; he could; on his
>
> BELLES: of his; (he was 1310); (he had 935); (in his 795); (that he 750); (I was 570); and his

The un-bracketed chains occur in the top 50 of either FICTION or BELLES. The bracketed chains (with their frequencies estimated per million words) occur in the top 50 of both sub-corpora. Frequencies are significantly lower in BELLES. For example,

FICTION	he was 2050	he had 1865	I was 950
BELLES	he was 1310	he had 935	I was 570

A χ^2-test on the original raw frequencies confirms a clear difference ($\chi^2 = 13.02$, df = 2, $p < 0.01$). LEARNED has no personal pronouns at all in the top 50 two-word chains.

A different distribution of content words also becomes more visible. The only two-word chains in the top 50 which contain content words are

> FICTION: he said; the door; she said
>
> BELLES: (the first); (the same)
>
> LEARNED: (the same); number of; (the first); for example; the other; such as

BELLES has no unique items, but the other items in LEARNED occur well outside the top 50 in BELLES: *the other* at rank 62, *such as* at 150, *number of* at 165, and *for example* at 232. The discourse markers which characterize LEARNED are already becoming visible.

4.3. THREE-WORD CHAINS

With three-word chains, constituent content words appear earlier on the lists and are more frequent, especially in LEARNED. The following are all the examples from the top 30:

> FICTION: in front of; I don't know; back to the; a couple of; side of the; for a moment; a lot of; was going to; the rest of; as soon as; (end of the); (the end of)
>
> BELLES: one of the; (as well as); (the end of); (the United States); (the fact that); (some of the); (part of the); (in order to); (on the other); at the time; at the same; (end of the); the same time; (a number of)
>
> LEARNED: (as well as); (a number of); (part of the); the number of; (the fact that); (some of the); in terms of; the use of; (in order to); (the end of); the presence of; (on the other); (the United States); (end of the); the case of; in the first; most of the; per cent of; in the case; the other hand; with respect to; the basis of

Again, it is BELLES which is not so clearly distinguished, and the frequencies of the chains shared by BELLES and LEARNED are not significantly different. Note that *the end of* and *end of the* occur in all three lists (see 9.2).

These lists illustrate an important characteristic of frequent phraseology. The term "content words" is misleading, since many of the words cited above occur in fixed phrases, where they do not carry their full lexical meaning (e.g. *number* in *a number of*). Data on the frequency of three-word chains illustrate how pervasive the delexicalization of frequent words is. In addition, frequent words are often frequent, not in their own right, but because they occur as part of frequent phrases (Summers 1996, Sinclair 1999). This means that the well known phenomenon of frequent words having more meanings than less frequent words (Zipf 1945) is not quite correctly phrased. For example, *rest* in *the rest of* clearly does not mean "relax, sleep".

One lexico-grammatical pattern is now evident. LEARNED has frequent chains with the pattern *DET N of*, where N is an abstract noun. All of the following occur in the top 50.

> LEARNED: a number of; the number of; the use of; the presence of; the case of; the basis of; the nature of; the development of; the effects of; a variety of; the rate of; the effect of [all over 90 per million]

None of these occur in the top 50 from FICTION, which has different *DET N of* chains.

> FICTION: a couple of; a lot of; the rest of; the end of; the back of; the edge of [all over 88 per million]

> BELLES has only two such chains *a matter of*, and *the end of*, which occurs in both other sub-corpora.

In addition, the sub-corpora are characterized by well known verbal and nominal styles. FICTION has several simple and compound verb forms, BELLES and LEARNED have fewer verb forms. LEARNED has no unique items containing verbs and is not distinguished clearly from BELLES. These are the chains in the top 30:

> FICTION: (there was a); (it was a); there was no; I don't know; it was the; (he had been); (that he was); (that he had); (he was a); it had been; (it would be); he did not; was going to; (that it was); to do with; would have been; he had to; to be a

> BELLES: (it is not); (it was a); to be a; (there is a); (that he was); (there was a); (that it was); it is a; (there is no); (that he had); (that it is); (he had been); (he was a); (it would be)

> LEARNED: (there is a); (it is not); (there is no); (that it is)

4.4. FOUR-WORD CHAINS

Four-word chains show clear differences in topic and function. In FICTION, the most frequent four-word chains are positional phrases with the patterns *the N of the* and *PREP the N of*. The following are all in the top 10:

> FICTION: (the rest of the); the edge of the; the end of the; in the middle of; in front of the; at the end of; the middle of the; the side of the [all over 28 per million]

Also frequent in the top 50 are combinations of a personal pronoun and one of a restricted number of verbs:

> FICTION: he was going to; I don't want to; I want you to; I don't know what; if you want to; you want me to; what do you think; are you going to; what do you mean; as if he were [all over 22 per million]

In BELLES there are only two verbs in the top 50 chains: *to be found in*; *seems to have been*. Neither BELLES nor LEARNED have any personal pronouns at all in the top 50 chains, but both have several meta-discourse markers, signalling the chronological or logical structure of the text. The following are all in the top 30:

> BELLES: (at the same time); (on the other hand); for the first time; (in the case of); (one of the most); (the rest of the); (on the basis of); at the time of; (as a result of); in the first place; on the part of; (as well as the); by the end of; on the one hand, in the face of; at the beginning of; in the history of; in the middle of [all over 26 per million]
>
> LEARNED: (on the other hand); (in the case of); (at the same time); (on the basis of); (as a result of); (as well as the); in terms of the; in the context of; the fact that the; as a function of; (one of the most); (in the form of); it is difficult to; it is clear that; it is necessary to; it is possible to [all over 38 per million]

The bracketed items discriminate between the text-types, though not very strongly. For example, the frequencies of two of the top chains (estimated per million words) are :

BELLES	on the other hand 90	in the case of 50
LEARNED	on the other hand 115	in the case of 110

A χ^2-test on the original raw frequencies confirms a difference at a modest level of significance ($\chi^2 = 3.97$, df = 1, $p < 0.05$).

4.5. FIVE-WORD CHAINS

Data on five-word chains add nothing entirely new, but some patterns become still clearer. In FICTION, there is only one meta-discourse marker in the top 10 (*as a matter of fact*), and eight out of the top 10 are position expressions:

> FICTION: (in the top 10) in the middle of the; the other side of the; (at the end of the); on the other side of; on the edge of the; the far end of the; at the far end of; on the side of the [all over 16 per million]

There are auxiliary verbs BE and HAVE, plus only a small set of main verbs:

> FICTION: (in the top 50) going, do, turned (out), opened, want, done, seemed, sat, looked, doing, do, got, thought, mean

In the top 10 in both BELLES and LEARNED, there are only two time/position expressions but several meta-discourse markers:

> BELLES and LEARNED: (at the end of the); at the beginning of the

> BELLES: (in the case of the); as a result of the; (on the part of the); as a matter of fact; for the first time in; (on the basis of the)

> LEARNED: (in the case of the); on the other hand the; (on the basis of the); (on the part of the); from the point of view

The top chain, *in the case of the*, is more frequent in LEARNED than BELLES (35 versus 20 per million), but the frequency of individual chains is becoming rather low here, and only the unbracketed chains could be recommended as discriminators.

5. EXTENDING THE METHOD WITH STOP-LISTS

Some chains discriminate less well because they occur in two (very rarely in all three) of the sub-corpora, but these chains can be excluded by setting up stop-lists much as we did for individual words (see 3.2 and 3.3). We constructed stop-lists of the top 100 two-, three-, four- and five-word chains in the combined four-million word corpus, and then deleted these chains from the top 100 chains in each sub-corpus. The resulting complete lists are in the Appendix. If this method was applied to a wider range of text-types, it could provide the basis of reference lists of diagnostic chains. Here we have space for just a few notes on some of the most striking characteristics of the resultant lists.

With two-and three-word chains, FICTION and LEARNED show clear patterns, but BELLES is less clear.

(a) Two-word chains.

> FICTION [40 items]: over half the chains contain a personal pronoun; there are only a few high frequency verbs (auxiliaries plus *said*, *want*, *know*)

> BELLES [15 items]: this is a rather short list with no very striking features; it has only a few auxiliary verbs (*does*, *is*, *had*, *was*)

> LEARNED [33 items]: there is only one pronoun (*we*); around a third of the chains are discourse markers and logical connectors (e.g. *for example*, *such as*, *at least*, *rather than*)

(b) Three-word chains.

> FICTION [55 items]: over a quarter of the chains are time and place expressions (e.g. *a long time, for a moment, down to the, the other side*)

> BELLES [21 items]: this is also a short list with no very striking features; it has several incomplete time expressions (e.g. *in the early*, *the age of*)

> LEARNED [53 items]: there are many logical expressions (e.g. *a result of*, *by means of*) and modality expressions (e.g. *is likely to, it is possible*); over a third of

> chains have the form *DET + abstract NOUN + of* (e.g. *a result of, the concept of*)

With four- and five-word chains, numbers of individual chains are becoming low, but lexico-grammatical patterns are clear.

(c) Four-word chains.

Numbers of individual chains (between 10 and 30) are becoming rather small here, and since each sub-corpus consisted of over 300 text samples, any specific chain can have occurred in only ten per cent of texts at most. In addition, some chains (e.g. BELLES *the first world war*) are obviously too topic-specific to act as general discriminators. However, the alphabetically ordered lists in the Appendix automatically bring together exponents of the lexico-grammatical patterns which characterize the sub-corpora, for example:

> FICTION [70 items]: at the edge of; at the foot of; etc; I don't know I; I don't know what; etc; the back of his; the direction of the; the front of the; etc.
>
> BELLES [54 items]: in the eyes of; in the field of; in the wake of; etc; the basis of the; the head of the; the life of the; etc.
>
> LEARNED [47 items]: in the absence of; in the presence of; in the range of; etc; the effect of the; the existence of a; the presence of a; the state of the; etc.

It is also revealing just to list the four top chains. FICTION has two position expressions, and two with a first person pronoun plus *want*; BELLES has time expressions; and LEARNED has logical expressions and discourse markers.

> FICTION: in front of him; on the edge of; I want you to; I don't want to [all over 32 per million]
>
> BELLES: of the nineteenth century; the turn of the; the history of the; the first world war [all over 26 per million]
>
> LEARNED: as a function of; in the presence of; with respect to the; it is important to [all over 35 per million]

(d) Five-word chains.

Here individual chains are not reliable discriminators: their frequencies are low (mainly between 5 and 15 per million), and each list is a mixture of chains which characterize the text-types and chains which are due to the topic of individual texts. Compare, (1) and (2) just for LEARNED:

> LEARNED: (1) it has been suggested that; it is assumed that the; it is important to note; etc; (2) a world in which Oswald; increase in the basic wage; of the central nervous system; etc.

The idiosyncratic chains in (2) are obviously useless as discriminators. However, the frequency of longer recurrent chains is itself a good discriminator, and we show in the next section that this provides an additional criterion for distinguishing text-types.

6. LONGER CHAINS AND REPETITIVE TEXT-TYPES

If turned into a *reductio ad absurdum*, then our claim that longer chains discriminate between text-types is self-evident, since if a chain is long enough it will uniquely identify an individual text. In other words, there is clearly a limit to the length of chains which recur in normal texts. For example, a forty-word chain would be unlikely to occur frequently in a single text, or to be repeated in independent texts by different authors (and if it was, it might well be evidence of plagiarism). However, our claims are rather more precise: text-types can be distinguished by chains of up to five words in length and also by the length of chains which recur in them.

In LEARNED, individual recurrent six-word chains still characterize the text-type, since they include several meta-discourse markers:

> *from the point of view of; in such a way as to; it is interesting to note that; it has been suggested that the* [all over 7 times per million]

By the time we reach seven-word chains, the only example of such a discourse marker to occur more than twice is

> *it is interesting to note that this*

Many other seven-word chains do recur, but are due to the specific topic of individual texts, and could not be used, individually, to make predictions about text-type:

> *a change in color equivalent to gray; a world in which Oswald did not; exhibition at the museum of modern art*

However, a diagnostic feature of LEARNED is the mere frequency of recurrent chains of this length. LEARNED is simply much more repetitive than either FICTION or BELLES, in the sense of containing a larger number of longer recurrent chains, since LEARNED texts place no premium on elegant stylistic variation (as Youmans 1990 shows using different methods). The simple frequency of seven-word chains (estimates per million words) therefore distinguishes clearly between LEARNED and the other two text-types: see Table 1. Here it is repetitiveness *per se* which is the discriminator. A statistical test is not needed to show the significant difference between LEARNED and the other two sub-corpora. The difference between FICTION and BELLES is perhaps in the expected direction (BELLES has less repetition and more elegant variation), but is not statistically significant.

Table 1 about here.

Certain text-types, such as formulaic liturgical language (Youmans 1990), are much more repetitive than others. For example, in the King James translation of the Bible (ca 852,600 running words), around 65 seven-word chains occur 20 times or more, and a further 260 occur 10 times or more. Examples (which many English speakers will recognize as well-known phrases) are:

> *and the Lord spake unto Moses saying; the word of the Lord came unto; shall know that I am the Lord; and it came to pass in the; and it shall come to pass in*

Many instances of these chains are parts of still longer recurrent chains, and since the Bible is not a single homogeneous text, many of these chains are concentrated in smaller highly repetitive sections of the whole. A small corpus of sermons also provided many recurrent seven-word chains, such as

> *died that we might be made whole; Jesus you gave your life for us; suffered and died that we might be*

Other repetitive text-types include political speeches (e.g. as recorded in Hansard), and some kinds of legal texts. Systematic generalizations about repetitive text-types are beyond the scope of this paper.

7. A NOTE ON TAG-CHAINS

In text analysis there is always a fine judgement involved in deciding the appropriate degree of descriptive delicacy. Ideal techniques would bring out patterns as clearly as possible, without leaving us open to the accusation that "you can prove anything with statistics". If the analysis is restricted to chains of individual word-forms, then we may lose generality, but if we ignore individual words in favour of grammatical categories, then we may obscure precisely those features which distinguish text-types. For example, we showed above that both LEARNED and FICTION are characterized by chains of *DET N of*. However, the chains contain different nouns, such as

> FICTION: a couple of; a lot of; the rest of
>
> LEARNED: a number of; the presence of; the development of

In addition, any grammatical tagging depends on the assumptions about appropriate categories which underlie the tag-set[6]. Nevertheless, bearing these caveats in mind, data on tag-chains allow some generalizations about grammatical patterns. For example, in the top 10 two-tag chains: only FICTION has either personal pronouns or past tense verbs, and they co-occur in chains such as *he was, he had*; only BELLES has proper nouns, in chains such as *Supreme Court*; and only LEARNED has noun-noun chains, such as *radio emission*, and verb past participles in chains such as *indicated by*, *observed in*.

Tag-chains do also reveal characteristics of English phraseology in general. The most frequent five-tag chains show that position expressions, with the pattern *PREP DET N PREP DET*, which we have already identified as frequent, are indeed amongst the most frequent chains in the language. Examples include: *on the end of the* (FICTION), *across the length of the* (BELLES) and *at the surface of the* (LEARNED). In 9.2 we therefore comment on this pattern.

8. SUMMARY OF METHODS AND FINDINGS

Our data add details to well-known studies which show that text-types are distinguished by lexical and grammatical patterns. FICTION is characterized by a verbal style, by past tense verb forms, and by frequent vocabulary from the lexical fields of saying, looking, thinking and wanting. LEARNED is characterized by a nominal style, by sub-technical vocabulary, and by a lack of stylistic variation. On several characteristics, BELLES texts lie between FICTION and LEARNED. However, previous studies document to only a very limited extent how multi-word chains and patterns of repetition contribute to text-type variation. We have shown that three distinct criteria can be used to discriminate between text-types:

1. the percentage of vocabulary from different normative lists (e.g. high frequency general and sub-technical words)

2. recurrent word-chains and/or their comparative frequency

3. measures of repetitiveness *per se* (which can distinguish LEARNED from the other two text-types).

For other studies which use related methods, see Milton (1998) and Aarts and Granger (1998) on using word-chains and tag-chains to distinguish text samples[7], and Youmans (1990), Altenberg (1998), Erman & Warren (2000) on alternative ways of measuring repetition in language use.

Word-frequency lists are a standard resource for many theoretical, descriptive and applied questions, but severe problems of definition mean that there are no equivalent phrase-frequency lists. Nevertheless, as we show, it is perfectly possible to investigate quantitative aspects of multi-word units. Comprehensive reference lists of chains of various lengths are well beyond the scope of this article, but if they were constructed from larger corpora and a wider range of text-types, such lists could provide norms which would have many applications in lexicography, in the preparation of language teaching materials, and in authorship attribution (stylistic and forensic).

9. INTERPRETATION

We now discuss some broader implications of our findings.

9.1. FROM THE OTHER END: ON CORPUS DESIGN

Text samples in corpora are often selected on external functional and sociolinguistic criteria, on the assumption that these samples are relatively homogeneous, but in some cases this assumption is doubtful, and the chains program can provide a quick check on homogeneity. For example, recurrent five-word chains in category D RELIGION in our four reference corpora suggest that the samples have been selected from different text-types, such as sermons, discussions of institutionalized religion, and philosophical/theological arguments:

> Brown: we are born of God; I say unto thee unless; the Lord is the strength
>
> LOB: in the Church of England; of the Dutch reformed church; the World Council of Churches
>
> Frown: that God does not exist; an argument of that sort; Catholic tradition in sexual ethics
>
> FLOB: the book of Common Prayer; literal reading of the text; you create your own reality

The reason for this heterogeneity is simple: "religion" is not a text-type, but a topic-based category. However, the reader might reasonably ask whether our sub-corpora are equally heterogeneous. This is unlikely, since FICTION, BELLES and LEARNED are text-types, but at least a rough check is simple to carry out. First, five-word chains from these text-type samples in the individual reference corpora do not differ in the way that the chains from RELIGION do. Second, a simple measure of repetitiveness gives confidence in the homogeneity of the samples. The number of five-word chains occurring twice (and, in brackets, three times or more) is shown in Table 2. The figures for each reference corpus are fairly consistent (a χ^2-test on the whole table is not statistically significant) and gives no reason to assume that the samples are from different populations.

Table 2 about here.

It is again clear (see section 6) that the LEARNED texts are much more repetitive, though it might be worth checking if the drop in repetitions from the 1961 texts of Brown and LOB, to the 1991 texts of Frown and FLOB is due to a shift to a less formal academic style across these thirty years: a comparison of Brown plus LOB versus Frown plus FLOB shows a statistically significant difference. Our methods complement similarity measures (such as type-token ratio and lexical density) which can characterize text-types on internal linguistic criteria, but detailed homogeneity and similarity measures are beyond the scope of this paper (for different possibilities, see Kilgarriff 2001). We started with sub-corpora designed on external criteria, and hypothesized that they have different internal features. Once these internal features have been confidently established, the argument can be reversed, and these internal criteria can be used as an automatic method of assessing the range of texts which constitute a corpus.

9.2. CHAINS AS EVIDENCE OF LINGUISTIC UNITS

So far, we have discussed recurrent chains as a statistical feature of language behaviour. In terms of their status in the mental lexicon, chains are only an intermediate representation (Rieger 2001: 171) and not necessarily linguistic or psycholinguistic units, as is evident in the following illustrative examples of recurrent five-word chains from FICTION. Some are not syntactic units, although they may contain one (e.g. *in and out of the*), or strongly predict a completion (e.g. *there was no point in*). Some are linguistic units but not necessarily pre-constructed (*they looked at each other*); and some, perhaps a surprisingly large number, are structural units which express a common meaning in a habitual and idiomatic way (e.g. *for a moment or two*, *the corner of his eye, you see what I mean*). Many of these last two sets are exponents of more abstract units (e.g. *it seemed to him/me that*). Although chains are not themselves cognitive units, they provide evidence of the inseparability of behaviour and competence. As Jones (2002: 21) puts it, corpus data can tap into the mental lexicons of hundreds or thousands of speakers. It is plausible that sequences which occur frequently in a corpus, across the language of many independent speakers, have a cognitive status, and that chains are surface evidence of psycholinguistic units which are exploited in producing and interpreting fluent language use. They are one type of evidence of abstract cognitive schemas.

We pointed out above that a chain such as *at the end* is not a good text-type discriminator, because it is simply very frequent in general English. For example, in the combined 4-million-word reference corpora, the chain *at the end of the* is twice as frequent as any other five-word chain, and the top 20 five-word chains include the following, listed here in descending frequency (all between 7 and 35 per million running words).

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
on the part of the
at the top of the
at the time of the

on the basis of the

Almost 30 per cent of the top 100 chains (types) have this pattern *PREP the N of the.* If we count closely related chains, with an indefinite article (*in the context of a*) or with embedded adjectives, such as *the ADJ N of the* (*the far side of the*), then, over 40 per cent of the top 100 chains have this pattern.

As Hunston and Francis (2000: 96) find, a given grammatical pattern occurs with a restricted lexical set; a few words in the set occur very frequently and other semantically related words occur more rarely. In this pattern, most of the nouns denote the outer limit or the middle of areas of space or periods of time, and some can be used both for physical places and time periods (e.g *at the beginning of the chapter / month*, *in the middle of the room / night*, *at the end of the pier / morning*). Several of the nouns can be either body parts or place terms: *back*, *bottom*, *foot*, *head*, *heart*, *lip*, *side*. These nouns all seem to have 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 (1995).

Although the 4-million word corpus consists of 2,000 diverse text samples, it is important to check that this pattern is not due to some idiosyncrasy of this corpus. Independent confirmation comes from Biber et al (1999: 1015) who identify *at the end of the* as the most frequent five-word "lexical bundle" in academic prose, and from Stubbs (2002) who compares the frequencies of the chains above with their frequencies in the 100-million word British National Corpus (90 million words of written and 10 million words of spoken English) and gets remarkably similar results. So, these chains are not an artefact of the 4-million-word corpus, but are frequent five-word chains in general English, though their frequencies differ in written and spoken genres.

Such patterns provide further evidence of the extensive delexicalization of common words. In chains such as *at the head of the* and *at the foot of the*, the nouns do not have their literal stand-alone meaning. Some nouns are used metaphorically, or can hardly be given their literal meaning at all, for example:

in the course of the
on the eve of the
in the face of the
in the heat of the
at the height of the
in the lap of the
in the light of the
on the spur of the
at the turn of the
in the wake of the

Light has some relation to the meaning "brightness" but the phrase means "taking into consideration some circumstances or information", and in several cases the phrase is delexicalized to a general expression of position in time or space: *in the course of the* =

"during"; *on the eve of the* = "just before"; *at the turn of the* = "just before or after"; *in the wake of the* = "just after". In addition, these chains predict other following nouns, in longer chains: *lap* predicts *gods*; *spur* predicts *moment*; *turn* is almost always followed by *century*; *heat* is most often followed by *moment*; *height* is usually followed by something unpleasant, such as *war* or *fighting*; *course* is usually followed by a word for a period of time (*day*, *evening*) or by a word which can be interpreted as a period of time (*negotiations*, *war*); *wake* is usually followed by a word for some sudden, often unpleasant, event which has surprised the speaker (*collapse*, *debacle*, *disaster*, *extraordinary success*).

9.3. ROUTINE LANGUAGE USE AND NORMS

Corpus studies may eventually be judged by how far they can contribute to the classic problem of social order. It is not well understood how the behaviour of many different speakers can become co-ordinated and focussed around the norms that we recognize as the idiomatic language use of a speech community, and a complete range of positions has been proposed with respect to the relation between *langue* (language system), *parole* (language behaviour, performance) and competence (individual knowledge). Saussure argued for a sharp difference in principle between *langue* and *parole*, but Firth dismissed the dualism as "a quite unnecessary nuisance" (Firth 1957: 2n, cf Halliday 1978: 38, 51, Sinclair 1991: 103), whereas other traditions accept the dualism but argue for the "interdependence of *langue* and *parole*" (Traugott & Heine eds 1991: 1). More recently, Halliday (1991: 34) has argued that instance and system are the same phenomenon seen from different points of view: just as the weather is the day-to-day realization of the climatic system. Similarly, Chomsky argues that performance is not even acceptable evidence for competence, whereas others argue that competence and performance are "merely two ends of a scale, rather than exclusive categories" (Keenan 1975: 148), or that statistics are useful in describing competence (Youmans 1990: 584).

Difficulties in conceptualizing how norms and conventions arise out of variable individual behaviour can also be seen in the metaphors which different scholars have used. Le Page and Tabouret-Keller (1985) discuss how idiosyncratic and diffuse language behaviour becomes "focussed" around norms. Hopper (1987), Hopper and Traugott (1993) and Carter and Sealey (2000) talk of grammatical constructions and social conventions "emerging" diachronically out of routine behaviour. Cameron (1997: 444) talks of "repeated acts [...] which congeal over time", and Teubert (1999) uses the rather similar phrase "semantic coagulation". Lenk (2000) talks of habitual patterns "stabilizing" through routine use. The use of such metaphors ("focussing", "emerging", "congealing", "coagulating" and "stabilizing") reveal conceptual problems, but they can be given an empirical basis in the data and methods we have illustrated.

Many studies have demonstrated that corpora and appropriate software make many phenomena visible for the first time, and our corpus data show how previously unobserved multi-word chains can make a modest contribution to understanding routine language use. In a Saussurian phrasing, we might ask how variable units of individual *parole* become focussed around the norms of social *langue*. This phrasing captures our position that it is useful to distinguish between *parole* and *langue*, but that corpus data throw doubt on the opposition in its traditional form and make it possible to study how frequent use leads to conventionalized structure. Sinclair (1991) makes two related statements which are often quoted. One is that pre-corpus linguistics was "starved of adequate data" (p.1), and the second is that "language looks rather different when you look at a lot of it at once" (p.100). The first statement is a criticism of the neo-Chomskyan reliance on small numbers of invented sentences. The second is a comment on observational technology: for example, the "chains" software makes visible patterns of repeated phraseology which are simply not visible without corpora and software. But there is a third point. When a lot of language is collected together, it not only *looks* different. "More *is* different" in the sense that whole systems exhibit behaviour which is qualitatively different from the behaviour of their individual lexical and syntactic constituents[8]. We have illustrated some features of text-types and of general English which are not predictable from the behaviour of their parts.

ACKNOWLEDGEMENTS

We are grateful to a seminar audience at the University of Stockholm in April 2003 and to three anonymous readers for helpful critical comments on an earlier version of this article.

NOTES

[1] Brown: one million words of written American English, published in 1961 (prepared under the direction of W. Nelson Francis and Henry Kucera); LOB: one million words of written British English, published in 1961 (prepared under the direction of Geoffrey Leech and Stig Johansson); Frown: one million words of written American English, published in 1991 (prepared under the direction of Christian Mair); FLOB: one million words of written British English, published in 1991 (prepared under the direction of Christian Mair).

[2] Category P, romance and love stories, was excluded merely to keep the sub-corpora of roughly equal size: FICTION is already larger than the other sub-corpora.

[3] That is, those words which are significantly more frequent in the sub-corpora than in a reference corpus. If the sub-corpora are compared with a reference corpus consisting of combined Brown, LOB, Frown and FLOB, the top 10 content keywords from FICTION contain several verbs (*said*, *looked*, *got*, *know*, *knew*), and hint at a preference for psychological topics (*looked*, *eyes*, *know*, *knew*); BELLES has only one verb (*wrote*), and vocabulary from artistic and literary topics (e.g. *art*, *literary*, *poetry*); and LEARNED has no verbs at all, and vocabulary from technical topics (e.g. *formula*, *data*, *system*). WordSmith Tools provides different statistics for comparing a sub-corpus and a reference corpus. The log-likelihood statistic was used here.

[4] The list, along with software which can compare texts and corpora against the word-list, is available from http://www.vuw.ac.nz/lals/software.htm.

5. We have used software written by Isabel Barth. Similar software has now been made available by William H. Fletcher at http://kwicfinder.com/kfNgram/.

6. We have used the Brill tagger, written by Eric Brill, and available at http://www.cs.jhu.edu.

7. It is often difficult to make direct comparisons between such studies. For example, Aarts and Granger (1998) identify three-tag chains in different corpora, and provide a small amount of normative data, but since they use a different tagger, it is not clear whether their chains are identical to ours.

8. "More is different" is the title of an article written by the Nobel prize-winning physicist Philip W. Anderson in 1972, and also the title of a collection of his articles (edited by N. P. Ong and R. N Blatt, Princeton University Press, 2001). The book is reviewed, with a useful discussion of the principle, in *The Times Higher Education Supplement*, 20 September 2002.

REFERENCES

Aarts, J. and Granger, S. (1998) Tag sequences in learner corpora. In S. Granger ed *Learner English on Computer*. London: Longman. 132-41.

Altenberg, B. (1998) On the phraseology of spoken English: the evidence of recurrent word combinations. In A. P. Cowie ed *Phraseology: Theory, Analysis and Applications*. Oxford: Oxford University Press. 101-122.

Bauer, L. and Nation, P. (1993) Word families. *International Journal of Lexicography*, 6, 4: 253-79.

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

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

Cameron, D. (1997) Performing gender identity. In A. Jaworski and N. Coupland eds *The Discourse Reader*. London: Routledge, 1999. 442-58.

Carter, B. and Sealey, A. (2000) Language, structure and agency. *Journal of Sociolinguistics*, 4, 1: 3-20.

Chomsky, N. (2000) *New Horizons in the Study of Language and Mind*. Cambridge: Cambridge University Press.

Cobuild (1995) *Collins COBUILD English Dictionary*. London: HarperCollins.

Coxhead, A. (2000) A new academic word list. *TESOL Quarterly*, 34, 2: 213-238.

Erman, B. and Warren, B. (2000). The idiom principle and the open choice principle. *Text*, 20, 1: 29-62.

Fillmore, C., Kay, P. and O'Connor, M. C. (1988) Regularity and idiomaticity in grammatical constructions. *Language*, 64: 501-38.

Firth, J. R. (1935) The technique of semantics. *Transactions of the Philological Society*. 36-72.

Firth, J. R. (1957) A synopsis of linguistic theory, 1930-1955. *Studies in Linguistic Analysis*. Special Volume, Philological Society. Oxford: Blackwell. 1-32.

Flowerdew, J.(1993) Concordancing as a tool in course design. *System*, 21, 2: 231-44. Also in M. Ghadessy, A. Henry & R. L. Roseberry eds *Small Corpus Studies and ELT*. Amsterdam: Benjamins, 2000. 71-92.

Halliday, M. A. K. (1978) *Language as Social Semiotic*. London: Edward Arnold.

Halliday M. A. K. (1991) Corpus studies and probabilistic grammar. In K. Aijmer and B. Altenberg eds *English Corpus Linguistics*. London: Longman. 30-43.

Hopper, P. (1987) Emergent grammar. In J. Aske, N. Berry, L. Michaelis and H. Filip eds *Berkeley Linguistics Society*, 13: 139-57.

Hopper, P. and Traugott, E. (1993) *Grammaticalization*. Cambridge: Cambridge University Press.

Hunston, S. and Francis, G. (2000) *Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English*. Amsterdam: Benjamins.

Johansson, S. (1981). Word frequencies in different types of English texts. *ICAME News*, 5: 1-13.

Jones, S. (2002) *Antonymy*. London: Routledge.

Keenan, E. L. (1975) Variation in universal grammar. In R. W. Fasold and R. W. Shuy eds *Analysing Variation in Language*. Washinton DC: Georgetown University Press. 136-48.

Kennedy, G. (1998) *An Introduction to Corpus Linguistics*. London: Longman.

Kilgarriff, A. (1997) Putting frequencies in the dictionary. *International Journal of Lexicography*, 10, 2: 135-55.

Kilgarriff, A. (2001) Comparing corpora. *International Journal of Corpus Linguistics*, 6, 1: 97-133.

Le Page, R. and Tabouret-Keller, A. (1985) *Acts of Identity*. Cambridge: Cambridge University Press.

Lenk, J. (2000) Stabilized expressions in spoken discourse. In C. Mair and M. Hundt eds *Corpus Linguistics and Linguistic Theory*. Amsterdam: Rodopi. 187-200.

Milton, J. (1998) Exploiting L1 and interlanguage corpora in the design of an electronic language learning and production environment.. In S. Granger ed *Learner English on Computer*. London: Longman. 186-98.

Nation, P. (2001) Using small corpora to investigate learner needs. In M. Ghadessy, A. Henry & R. L. Roseberry eds *Small Corpus Studies and ELT*. Amsterdam: Benjamins. 31-45.

Pawley, A. and Syder, F. H. 1983. Two puzzles for linguistic theory. In J. C. Richards and R. W. Schmidt eds *Language and Communication*. London: Longman. 191-226.

Piotrowski, R. G. (1984) *Text, Computer, Mensch*. Bochum: Brockmeyer.

Rieger, B. (2001) Computing granular word meanings. In P. Wang ed *Computing with Words*. NY: Wiley. 147-208.

Scott, M. (1997a) *WordSmith Tools Manual*. Oxford: Oxford University Press.

Scott, M. (1997b). PC analysis of keywords, and key key words. *System*, 25, 2: 233-45.

Sinclair, J. (1991) *Corpus Concordance Collocation*. Oxford: Oxford University Press.

Sinclair, J. (1999) A way with common words. In H. Hasselgard and S. Oksefjell eds *Out of Corpora*. Amsterdam: Rodopi. 157-79.

Strzalkowski, T. ed (1998) *Natural Language Information Retrieval*. Dordrecht: Kluwer.

Stubbs, M. (2001) On inference theories and code theories: corpus evidence for semantic schemas. *Text*, 21, 3: 437-65.

Stubbs, M. (2002) Two quantitative methods of studying phraseology in English. *International Journal of Corpus Linguistics*, 7, 2: 215-44.

Summers, D. (1996) Computer lexicography: the importance of representativeness in relation to frequency. In J. Thomas and M. Short eds *Using Corpora for Language Research*. London: Longman. 260-66.

Teubert, W. (1999). Korpuslinguistik und Lexikographie. *Deutsche Sprache*, 4/1999, 292-313.

Traugott, E. and Heine, B. eds (1991) *Approaches to Grammaticalization*. 2 volumes. Amsterdam: Benjamins.

West, M. (1953). *A General Service List of English Words*. London: Longman.

Yang, H. (1986) A new technique for identifying scientific / technical terms and describing science texts. *Literary and Linguistic Computing*, 1, 2: 93-103.

Youmans, G. (1990). Measuring lexical style and competence: the type-token vocabulary curve. *Style*, 24, 4: 584-99.

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

Zipf, G. K. (1945). The meaning-frequency relationship of words. *Journal of General Psychology*, 33: 251-56.

Table 1. Frequency of seven-word chains in three subcorpora.

	seven-word chains occurring 3 times or more	seven-word chains occurring twice
FICTION:	10	400
BELLES:	5	280
LEARNED:	150	1,960

Table 2. Frequency of five-word chains in four reference corpora.

	FICTION	BELLES	LEARNED
Brown 1961:	360 (51)	292 (31)	1212 (198)
LOB 1961:	376 (44)	310 (35)	1136 (180)
Frown 1991:	378 (37)	337 (39)	1122 (131)
FLOB 1991 :	348 (35)	302 (43)	961 (117)

APPENDIX

These lists have been compiled in the following way. The 100 most frequent two-, three-, four- and five-word chains were extracted from the combined four-million word corpus, placed in stop-lists, and then deleted from the 100 most frequent chains of each length in each sub-corpus. So, the lists below are the top 100 chains in each sub-corpus, minus the chains in the stop-lists. The resulting lists are obviously of different lengths: a longer list reveals that the sub-corpus has more distinctive chains: that is, chains which are not also high in frequency in the combined corpus. The lists have been alphabetized, which makes it easier to see chains which are similar in their syntax and semantics.

FICTION: top two-word chains (minus stop-list).

332,	a few
359,	a little
408,	and then
466,	as he
319,	as if
324,	back to
326,	but he
289,	but I
310,	do you
454,	going to
308,	had a
366,	had to
420,	he could
308,	he would
275,	his head
291,	I could
458,	I don't
421,	I had
290,	I said
284,	if he
280,	if you
289,	in her
423,	like a
301,	of them
404,	on his
404,	she said
290,	that I
287,	that she
350,	that was
410,	the door
278,	the man
293,	they had
288,	to go
294,	to her
294,	to him
317,	to his
347,	to see
325,	up the
275,	want to
276,	you know

BELLES: top two-word chains (minus stop-list).

193,	a new
193,	and that

225,	and to
193,	does not
197,	had a
213,	he is
260,	his own
238,	I had
197,	in their
202,	is that
197,	is to
193,	of an
253,	to his
196,	was to
200,	who had

LEARNED: top two-word chains (minus stop-list).

195,	and that
216,	and to
186,	are not
185,	as an
193,	at a
218,	at least
290,	between the
228,	could be
231,	do not
273,	does not
186,	due to
205,	et al
323,	for example
190,	from a
304,	if the
243,	is that
245,	is to
185,	not be
255,	of an
251,	per cent
206,	rather than
220,	so that
194,	such a
307,	such as
191,	than the
206,	the second
259,	the two
246,	use of
244,	we have
190,	well as
265,	which is
286,	which the

```
202,  within the
```

FICTION, top three-word chains (minus stop-list).

```
 65,  a long time
 63,  and he was
 54,  and there was
 53,  as if he
 49,  away from the
 51,  back into the
 55,  but he was
 59,  but there was
 60,  do you think
 52,  down on the
 50,  down to the
 49,  edge of the
106,  for a moment
 62,  front of the
 58,  going to be
 68,  had been a
 49,  he could not
 52,  he could see
 60,  he had a
 83,  he had to
 49,  he said I
 58,  he wanted to
 56,  he was not
 66,  I don't think
 52,  I had to
 55,  I have to
 54,  I want to
 54,  in the middle
 58,  in the morning
104,  it had been
 62,  I'm going to
 59,  might have been
 50,  on the floor
 49,  on to the
 54,  one of those
 66,  seemed to be
 52,  she had been
 55,  shook his head
 63,  that she had
 54,  that she was
 50,  that was the
 76,  the back of
 52,  the door and
 62,  the edge of
 53,  the middle of
 52,  the other side
 59,  the side of
 50,  the sound of
 50,  there had been
 54,  to the door
 92,  was going to
 58,  went to the
 65,  what do you
 55,  when he was
 61,  you want to
```

BELLES: top three-word chains (minus stop-list).

```
 49,  a kind of
 38,  have been a
 42,  I do not
 51,  in new york
 49,  in spite of
 37,  in the early
 37,  in the last
 38,  in which he
 42,  is that the
 41,  may have been
 38,  member of the
 37,  much of the
 37,  of his life
 56,  of his own
 40,  of the first
 41,  one of his
 40,  seems to have
 40,  so far as
 41,  such as the
 46,  that of the
 48,  the age of
 52,  the history of
 38,  the house of
 38,  the idea of
 37,  the kind of
 37,  the nineteenth century
 42,  the problem of
 42,  the time of
 37,  to do so
 43,  was to be
 46,  when he was
```

LEARNED: top three-word chains (minus stop-list).

44,	1 and 2
50,	a function of
44,	a result of
62,	a variety of
44,	according to the
49,	and that the
46,	as a whole
45,	as in the
44,	as shown in
46,	based on the
51,	be used to
44,	by means of
58,	due to the
62,	each of the
47,	for example the
47,	found in the
59,	in relation to
44,	in this case
51,	is likely to
47,	is that the
45,	it can be
45,	it is possible
61,	it should be
62,	likely to be
58,	nature of the
45,	of the first
50,	of the same
48,	of the two
50,	on the basis
50,	seems to be
73,	so that the
51,	such as the
57,	that of the
53,	the amount of
44,	the concept of
60,	the effect of
64,	the effects of
46,	the existence of
53,	the form of
53,	the importance of
55,	the level of
67,	the nature of
44,	the order of
105,	the presence of
47,	the problem of
51,	the range of

62,	the rate of
48,	the result of
49,	the role of
51,	the value of
51,	use of the
54,	used in the
72,	with respect to

FICTION: top four-word chains (minus stop-list).

13,	a matter of fact
13,	and at the same
19,	and there was a
15,	and there was no
15,	anything to do with
17,	are you going to
15,	as if he were
12,	as soon as he
12,	as soon as the
12,	at the edge of
19,	at the far end
14,	at the foot of
13,	at the side of
22,	but there was no
14,	by the time he
17,	far end of the
12,	get out of here
14,	he could see the
12,	he didn't want to
14,	he looked at her
13,	he shook his head
15,	he went to the
12,	he would have to
12,	I don't know I
22,	I don't know what
12,	I don't think I
23,	I don't want to
23,	I want you to
14,	if it had been
22,	if you want to
27,	in front of him
12,	in the direction of
16,	it had been a
15,	it might have been
12,	it must have been
17,	it was as if
12,	it was the only
19,	on the back of

```
 23,  on the edge of
 13,  on the side of
 13,  on top of the
 12,  out of the car
 15,  out of the house
 12,  out of the way
 12,  sat down on the
 15,  she shook her head
 14,  she was going to
 13,  side of the road
 12,  something to do with
 15,  that he had been
 14,  that she had been
 20,  the back of his
 15,  the direction of the
 17,  the far end of
 13,  the foot of the
 19,  the front of the
 12,  the side of his
 19,  the two of them
 15,  there had been no
 12,  there was no way
 17,  there was only one
 14,  to get out of
 13,  to go to the
 19,  was going to be
 14,  went back to the
 17,  what do you mean
 18,  what do you think
 15,  what was going on
 13,  when I was a
 20,  you want me to
```

BELLES: top four-word chains (minus stop-list).

```
 13,  a good deal of
 10,  a matter of fact
 13,  a part of the
 14,  as one of the
 11,  as part of the
 10,  as well as a
 14,  at a time when
 12,  at the level of
 11,  he did not want
 10,  I do not think
 12,  in so far as
 12,  in the eighteenth century
 10,  in the eyes of
```

15,	in the field of
14,	in the nineteenth century
15,	in the spring of
10,	in the summer of
10,	in the wake of
10,	it is not surprising
10,	it is not the
12,	it is true that
10,	it must have been
11,	it seems to me
11,	of the eighteenth century
14,	of the history of
19,	of the nineteenth century
12,	one of the best
14,	one of the few
15,	seems to have been
12,	so far as to
10,	that he had been
11,	that it is not
10,	that it was a
11,	that it would be
11,	that there was a
11,	the basis of the
13,	the development of the
11,	the fact that he
10,	the first half of
16,	the first world war
11,	the head of the
16,	the history of the
13,	the house of commons
15,	the house of lords
10,	the life of the
10,	the name of the
12,	the new york times
13,	the second world war
18,	the turn of the
10,	this is not to
10,	to have been a
10,	to say that the
11,	to the point of
15,	turn of the century

LEARNED: top four-word chains (minus stop-list).

19,	a function of the
18,	a great deal of
15,	a high degree of
15,	a world in which

15, are more likely to
29, as a function of
14, as shown in fig
16, as we have seen
16, as we shall see
15, at the present time
15, be found in the
14, by means of a
19, can be used to
16, in a number of
16, in contrast to the
17, in each of the
18, in relation to the
13, in so far as
21, in the absence of
13, in the development of
15, in the number of
28, in the presence of
14, in the range of
16, in the sense that
16, is based on the
13, is the fact that
23, it is important to
20, of a number of
19, of the order of
15, the basic wage rate
17, the basis of the
14, the degree to which
15, the effect of the
13, the effects of noise
14, the existence of a
14, the plane of the
13, the presence of a
14, the second half of
21, the state of the
13, the strength of the
16, the structure of the
22, the surface of the
20, the total number of
13, the validity of the
13, there is only one
17, to the extent that
27, with respect to the

FICTION: top five-word chains (minus stop-list).

4, a fat hell on big
4, all I could think of

```
 4,  all I know is that
 4,  am I going to do
13,  and at the same time
 4,  and by the time I
 4,  and have a look at
 4,  and looked up at the
 4,  and walked over to the
 5,  anything to do with the
 8,  are you going to do
 9,  as if it had been
 4,  as though it were a
 6,  at the edge of the
 5,  at the front of the
 8,  at the side of the
 4,  at the top of her
 4,  both sides of the road
 5,  by the time he was
 4,  do you want me to
 4,  done this before they always
 4,  down the front of his
 4,  fat hell on big hans
 4,  for a moment or two
 4,  get the hell out of
 5,  got out of the car
 4,  had a right to know
 5,  had anything to do with
 5,  had been one of the
 4,  had not been able to
 4,  had something to do with
 4,  have anything to do with
 4,  he felt as if he
 4,  he got out of the
 5,  he had no time to
 4,  he knew that he had
 4,  he looked at his watch
 6,  he opened the door and
 4,  he won't be able to
 6,  I don't want to be
 5,  I thought it was a
 4,  I want you to be
 5,  if he was going to
 5,  if it hadn't been for
 4,  if you want to know
 4,  in a place like this
 5,  in and out of the
 5,  in front of him he
 4,  in front of the mirror
 5,  in the back of the
```

```
4,  in the back seat of
5,  in the shadow of the
4,  it for a long time
6,  it seemed to him that
4,  it seemed to me that
4,  it was a long time
4,  it was a relief to
4,  it was easy to see
4,  it was one of the
4,  it was one of those
5,  it was the only thing
6,  it was too late to
6,  I've never done this before
5,  looked as if it had
4,  made his way to the
7,  mr ball and mr baring
4,  never done this before they
4,  on both sides of the
6,  on the back of his
4,  on the far side of
4,  on the way to the
4,  opened the door to the
4,  out of the car and
4,  out of the corner of
4,  put his arm around her
4,  sat back in his chair
6,  she sat down on the
5,  the back of his hand
4,  the back of his head
4,  the back of the house
5,  the corner of his eye
4,  the edge of the bed
5,  the end of the bar
4,  the end of the street
4,  the far side of the
4,  the first time he had
4,  the foot of the bed
5,  the foot of the stairs
4,  the friends of eddie coyle
5,  the front door of the
4,  the middle of the road
5,  the middle of the room
4,  the open door of the
6,  the opposite side of the
4,  the rest of the evening
4,  the side of his head
6,  the side of the road
5,  the top of the stairs
```

```
  4,  the upper part of his
  4,  the way back to the
  4,  there are a lot of
  4,  there was a bit of
  5,  there was no point in
  4,  there was no reason why
  8,  there was no sign of
  4,  there was of course no
  5,  there would be time for
  6,  they looked at each other
  4,  this before they always said
  4,  to get out of the
  4,  to have a word with
  9,  to the edge of the
  5,  to the front of the
  4,  to the other side of
  4,  to the side of the
  4,  was a bit of a
  5,  was going to be a
  4,  what am I going to
  6,  what are you doing here
  7,  what are you going to
  5,  what he was doing he
  4,  what is it he asked
  4,  what I'm going to do
  5,  with the back of his
  5,  you see what I mean
  5,  you want me to do
```

BELLES: top five-word chains (minus stop-list).

```
  4,  a few years ago in
  4,  a matter of fact the
  4,  a quarter of a century
  5,  a year and a half
  4,  all men are created equal
  4,  and a host of other
  4,  as a body of knowledge
  4,  as I have already said
  4,  as well as on the
  4,  at the end of july
  6,  at the head of the
  4,  at the heart of the
  4,  at the level of the
  6,  at the same time he
  4,  at the time of writing
  4,  be an integral part of
  4,  book and the war are
```

```
4,  contact with the outside world
4,  during the next five years
5,  during the second world war
5,  has to do with the
7,  he did not want to
4,  I do not think that
6,  in one way or another
4,  in the development of the
5,  in the first place the
4,  in the history of the
4,  in the midst of the
4,  it is as if the
4,  it is not easy to
4,  it may be that the
4,  it must be admitted that
5,  it was one of the
7,  madonna of the rose garden
4,  my book and the war
5,  nothing to do with the
4,  of the first world war
4,  of the kind and unkind
4,  of the new york times
4,  of the second world war
4,  of the way in which
4,  of the wealth of nations
4,  on the face of it
4,  on the other hand if
4,  on the west side of
5,  sheffield trades and labour council
4,  that he would not be
6,  the eighteenth and nineteenth centuries
4,  the fact that he was
5,  the house of lords the
5,  the madonna of the rose
4,  the rest of his life
5,  the sheffield trades and labour
4,  the walnut trees of altenburg
4,  the way in which the
5,  the way to the churchyard
4,  them what do you think
5,  there is no evidence that
4,  to be an integral part
4,  to come to terms with
4,  to take advantage of the
5,  turned out to be the
4,  was a close friend of
```

LEARNED: top five-word chains (minus stop-list).

```
 8,  a change in color equivalent
 5,  a large part of the
 5,  a topic for overt discussion
 5,  a world in which oswald
 5,  an index word or electronic
 5,  an integral function of l
 9,  as a function of the
 6,  as a function of time
 6,  as can be seen from
 6,  at the same time to
 8,  change in color equivalent to
 8,  color equivalent to gray scale
 6,  does not seem to be
 8,  equivalent to gray scale step
 5,  for which formula for every
 5,  formula for every state alpha
 5,  has been suggested that the
 5,  in a wide range of
 8,  in color equivalent to gray
 6,  in each of the four
11,  in the basic wage rate
 6,  in the course of his
 5,  in the first part of
 5,  in the region of the
 5,  in the sense that the
 7,  in the vicinity of the
 5,  increase in the basic wage
11,  index word or electronic switch
 8,  index words and electronic switches
 6,  is interesting to note that
 5,  is the fact that the
10,  it can be seen that
 7,  it has been suggested that
 6,  it is assumed that the
 5,  it is important to note
 7,  it is interesting to note
 8,  it should be possible to
 6,  it should be remembered that
 5,  it was found that the
 6,  it will be seen that
 5,  may or may not be
 5,  nomenclature symbols units physical constants:_
 5,  number of terms previous schooling
 5,  of nomenclature symbols units physical
 5,  of the central nervous system
 5,  of the game of chess
 5,  of the plan must be
 5,  of the united states code
 8,  on the basis of a
 6,  on the gray scale for
 5,  on the other hand if
 5,  on the other hand it
 6,  on the other hand there
 7,  on the surface of the
 5,  principle of the plan must
 6,  rate of change of the
 6,  refund of taxes paid by
 7,  that it is difficult to
10,  the effects of noise on
 5,  the eigenfunctions and adjoint functions
 6,  the extent to which the
 5,  the federal rules of evidence
 6,  the first part of the
 8,  the gray scale for staining
 6,  the iliad and the odyssey
 9,  the influence line for the
 5,  the level of excess demand
 6,  the minimal polynomial for ~t
 5,  the museum of modern art
 6,  the null space of f
 5,  the number of terms previous
 5,  the plan must be that
 5,  the plane of the pencil
 5,  the purpose of this paper
 5,  the radio emission of the
 8,  the rate of change of
 6,  the rest of the line
 5,  the state of the art
 5,  the state of the stream
 5,  the table of dictionary usage
 8,  the timing of an interpretation
 5,  the upper part of the
 5,  the way in which the
 6,  to the extent that the
 6,  to the fact that the
 5,  upregulation of the camp system
 5,  use of nomenclature symbols units
 5,  world trade in primary products
 8,  - a change in color
 6,  - staining equivalent to row
```