In A Davies & C Elder eds (2004) *Handbook of Applied Linguistics*. Oxford: Blackwell. 106-32.

LANGUAGE CORPORA

Michael Stubbs

Since the 1990s, a 'language corpus' usually means a text collection which is:

- large: millions, or even hundreds of millions, of running words, usually sampled from hundreds or thousands of individual texts
- computer-readable: accessible with software such as concordancers, which can find, list and sort linguistic patterns
- designed for linguistic analysis: selected according to a sociolinguistic theory of language variation, to provide a sample of specific text-types or a broad and balanced sample of a language.

Much 'corpus linguistics' is driven purely by curiosity. It aims to improve language description and theory, and the task for applied linguistics is to assess the relevance of this work to practical applications. Corpus data are essential for accurately describing language use, and have shown how lexis, grammar and semantics interact. This in turn has applications in language teaching, translation, forensic linguistics, and broader cultural analysis. In limited cases, applications can be direct. For example, if advanced language learners have access to a corpus, they can study for themselves how a word or grammatical construction is typically used in authentic data. Hunston (2002: 170-84) discusses data-driven discovery learning and gives further references.

However, applications are usually indirect. Corpora provide observable evidence about language use, which leads to new descriptions, which in turn are embodied in dictionaries, grammars and teaching materials. Since the late 1980s, the influence of this work is most evident in new monolingual English dictionaries (CIDE 1995, COBUILD 1995a, LDOCE 1995, OALD 1995) and grammars (e.g. COBUILD 1990), aimed at advanced learners, and based on authentic examples of current usage from large corpora. Other corpus-based reference grammars (e.g. G. Francis et al 1996, 1998, Biber et al 1999) are invaluable resources for materials producers and teachers.

Corpora are just sources of evidence, available to all linguists, theoretical or applied. A sociolinguist might use a corpus of audio-recorded conversations to study relations between social class and accent; a psycholinguist might use the same corpus to study slips of the tongue; and a lexicographer might be interested in the frequency of different phrases. The study might be purely descriptive: a grammarian might want to know which constructions are frequent in casual spoken language but rare in formal written language. Or it might have practical

aims: someone writing teaching materials might use a specialized corpus to discover which grammatical constructions occur in academic research articles; and a forensic linguist might want to study norms of language use, in order to estimate the likelihood that linguistic patterns in an anonymous letter are evidence of authorship.

So, if corpus linguistics is not (necessarily) applied linguistics, and is not a branch of linguistics, then what is it? It is an empirical approach to studying language, which uses observations of attested data in order to make generalizations about lexis, grammar and semantics. Corpora solve the problem of observing patterns of language use. It is these patterns which are the real object of study, and it is findings about recurrent lexico-grammatical units of meaning which have implications for both theoretical and applied linguistics. Large corpora have provided many new facts about words, phrases, grammar and meaning, even for English, which many teachers and linguists assumed was fairly well understood.

Valid applications of corpus studies depend on the design of corpora, the observational methods of analysis, and the interpretation of the findings. Applied linguists must assess this progression from evidence to interpretation to applications, and the article therefore has sections on empirical linguistics (pre- and post-computers), corpus design and software, findings and descriptions, and implications and applications.

I use these presentation conventions. LEMMAS (LEXEMES) are in upper case. *Word-forms* are lower case italics. "Meanings" are in double quotes. Collocates of a node are in diamond brackets: UNDERGO <surgery>.

1. EMPIRICAL LINGUISTICS

Since corpus study gives priority to observing millions of running words, computer technology is essential. This makes linguistics analogous to the natural sciences, where it is observational and measuring instruments (such as microscopes, radio telescopes and x-ray machines) which extended our grasp of reality far beyond 'the tiny sphere attainable by unaided common sense' (Wilson 1998, p.49).

Observation is not restricted to any single method, but concordances are essential for studying lexical, grammatical and semantic patterns. Printed concordance lines (see Appendix) are limited in being static, but a computer-accessible concordance is both an observational and experimental tool, since ordering it alphabetically to left and right brings together repeated lexico-grammatical patterns. A single concordance line, on the horizontal axis, is a fragment of language use (*parole*). The vertical axis of a concordance shows repeated co-occurrences, which are evidence of units of meaning in the language system (*langue*).

The tiny sample of concordance lines in the Appendix is not representative. In a real study one might have hundreds or thousands of concordance lines, but I can

use this sample as illustrations. Concordance data are often especially good at distinguishing words with related propositional meanings, but different connotations and patterns of usage. The Appendix therefore gives examples of *endure*, *persevere*, *persist* and *undergo*, which are all used to talk about unpleasant things which last a long time, but which differ in their surrounding lexis and grammar. For example, we can observe how the word-form *persist* occurs in distinct constructions. When its subject is an abstract noun, it often denotes unpleasant things (*fears*, *problems*), often medical (*symptoms*, *headaches*), and often has a time reference (*for over a year*, *for up to six weeks*). Alternatively, when the subject of *persist in* is animate, it is often used of someone who persists, often unreasonably or *in the face of* opposition, in doing something which is difficult or disapproved of. Such recurrent co-occurrence patterns provide evidence of typical meaning and use.

It is sometimes objected that concordances place words in small, arbitrary contexts, defined by the width of a computer screen, and ignore contexts of communication. However, it is an empirical finding that evidence for the meaning of a node word often occurs within a short span of co-text. In addition, corpora allow individual utterances to be interpreted against the usage of many speakers and the intertextual norms of general language use.

The observation of large publicly available data sets implies (a weak sense of) inductive methods, that is, gathering many observations and identifying patterns in them. This does not imply mechanical methods of generalizing from observations, but (as Fillmore 1992, pp. 38, 58 puts it) a combination of corpus linguistics (getting the facts right) and armchair linguistics (thinking through the hypotheses that corpus data suggest). It does mean, however, that corpus study belongs to a philosophical tradition of empiricism. Contrary to a loss of confidence, from Saussure to Chomsky, in the ability to observe real language events, corpora show that language use is highly patterned. Although there are limitations on corpus design (see below), and although we can never entirely escape subjective interpretations, corpora allow 'a degree of objectivity' about some central questions, 'where before we could only speculate' (Kilgarriff 1997, p.137). There are no automatic discovery procedures, but inductive generalizations can be tested against observations in independent corpora.

Corpus methods therefore differ sharply from the view, widely held since the 1960s, that native speaker introspection gives special access to linguistic competence. Although linguists' careful analyses of their own idiolects have revealed much about language and cognition, there are several problems with intuitive data and misunderstandings about the relation between observation and intuition in corpus work. Intuitive data can be circular: data and theory have the same source in the linguist who both proposes a hypothesis and invents examples to support or refute it. They can be unreliable or absent: many facts about frequency, grammar and meaning are systematic and evident in corpora, but unrecorded in pre-corpus dictionaries. They are narrow: introspection about small sets of invented sentences cannot be the sole and privileged source of data.

There is no point in being purist about data, and it is always advisable to compare data from different sources, both independent corpora, and also introspection and experiments. Corpus study does not reject intuition, but gives it a different role. Concordances focus intuition, and this 'confirms rather than produces the data' (De Beaugrande 1999, pp.247-48). Without this retrospective competence, native speakers could not recognize untypical collocations in literature, advertising or jokes. We cannot know in advance what kinds of evidence might bear on a theory of linguistic competence (as even Chomsky 2000: 139-40 admits). Nevertheless, with some striking exceptions (Fillmore 1992), cognitive approaches have neglected corpus data on recurrent semantic patterns as evidence of cognitive structures.

2. SOME BRIEF HISTORY

There was corpus study long before computers (W. Francis 1992) and, from a historical perspective, Saussure's radical uncertainty about the viability of studying *parole*, followed by Chomsky's reliance on introspective data, were short breaks in a long tradition of observational language study. Disregard of quantified textual data was never, of course, accepted by everyone. Corder (1973, pp.208-23) emphasizes the relevance of frequency studies to language teaching, and language corpora have always been indispensable in studying dead languages, unwritten languages and dialects, child language acquisition, and lexicography. So, within both philological and fieldwork traditions, corpus study goes back hundreds of years, within a broad tradition of rhetorical and textual analysis.

Early concordances were prepared of texts of cultural significance, such as the Bible (Cruden 1737). Ayscough's (1790) index of Shakespeare is designed 'to point out the different meanings to which words are applied'. Nowadays we would say that he had a concept of 'meaning as use'. By bringing together many instances of a word, a concordance provides evidence of its range of uses and therefore of its meanings, and this essential point is still the basis of corpus semantics today.

The other main reason for studying large text collections, which again emphasizes the central concern with meaning, was the attempt to produce comprehensive dictionaries. From Samuel Johnson's dictionary of 1755 onwards, lexicographers have used quotations to illustrate the uses and meanings of words. Johnson collected 150,000 illustrative quotations for 40,000 head-words, and the readers for the *Oxford English Dictionary* collected five million quotations to illustrate over 400,000 entries. (Kennedy 1998, pp.14-15, Winchester 1998.) For example, Johnson's dictionary has these quotes which contain *persist*:

- ... I would advise neither to *persist* in refusing
- ... the sinful act, to continue and *persist* in it
- ... thus will *persist*, relentless in his ire

The collocates of *persist* are observable evidence of its typical semantic features of doing something over time and against opposition. However, there is a limitation here on printed dictionaries: these examples do not occur under the head-word PERSIST, and can therefore be found only by a full text search of a machine-readable version of the dictionary. The Appendix gives further illustrations of observable evidence of meaning. For example, *endure* co-occurs with *compelled* and *forced*, *difficult* and *painful*, with references to long time periods, and also with near synonyms such as *persevere*, *accept* and *bear*. Semantic features are not abstract, but often realized in co-occurring and observable collocates.

Modern lexicographers use better designed corpora, their methods are more explicit, they use statistical techniques to systematize observations (Church & Hanks 1990, Clear 1993, Sinclair et al 1998), and the theory of 'meaning as use' has been developed by Wittgenstein, Austin and Firth, but the basic approach to semantic analysis is not fundamentally different from that of Cruden, Ayscough, Johnson and Murray.

Other impressive quantitative corpus analyses, between the 1890s and the 1950s, were possible only with significant expense and personnel, and often had precise institutional and/or educational applications. In order to improve shorthand methods for court transcription, Kaeding (1898) used large numbers of helpers from the Prussian civil service to analyse word frequency in an 11-million word German corpus. From the 1920s to the 1940s, Thorndike and Lorge (1944) calculated word frequencies in large English-language corpora, of up to 18-million words. These word-lists were used to control the vocabulary in foreign language and literacy materials. West's (1953) influential General Service List gave also the frequency of different meanings of words.

In a word, corpus-based study of language is much older than its alternatives. Indeed, up until the 1950s, it was assumed that writing a grammar required the study of text collections. Famous examples include: Jespersen (1909-49), based on examples of written English over several centuries; Fries (1952), based on a 250,000-word corpus of telephone conversations; and Quirk et al (1972), based on the last of the great non-computerized corpora, which was itself over-taken by technology and computerized, and then used in turn for later versions of the grammar (Quirk et al 1985, and, with substantial additional corpora, Biber et al 1999).

3. MODERN CORPORA AND SOFTWARE

Modern computer-assisted corpus study is based on two principles.

- The observer must not influence what is observed.

What is selected for observation depends on convenience, interests and hypotheses, but corpus data are part of natural language use, and not produced for purposes of linguistic analysis.

- Repeated events are significant.

Quantitative work with large corpora reveals what is central and typical, normal and expected. It follows (Teubert 1999) that corpus study is inherently sociolinguistic, since the data are authentic acts of communication; inherently diachronic, since the data are what has frequently occurred in the past; and inherently quantitative. This disposes of the frequent confusion that corpus study is concerned with 'mere' performance, in Chomsky's (1965, p.3) pejorative sense of being characterized by 'memory limitations, distractions, shifts of attention and interest, and errors'. The aim is not to study idiosyncratic details of performance which are, by chance, recorded in a corpus. On the contrary, a corpus reveals what frequently recurs, sometimes hundreds or thousands of times, and cannot possibly be due to chance.

AVAILABLE CORPORA

Any list of extant corpora would be quickly out of date, but there are two sets of important distinctions between

- small first generation corpora from the 1960s onwards and much larger corpora from the 1990s, and
- carefully designed reference corpora, small and large, and other specialized corpora, opportunistic text collections, archives and the like.

The first computer-readable corpora, in the 1960s, are very small by contemporary standards, but still useful because of their careful design. The Brown corpus (from Brown University in the USA) is one million words of written American English, sampled from texts published in 1961: both informative prose, from different text-types (e.g. press and academic writing), and different topics (e.g. religion and hobbies); and imaginative prose (e.g. detective fiction and romance). Parallel corpora were designed to enable comparative research: the LOB corpus (from the universities of Lancaster, Oslo and Bergen) contains British data from 1961; Frown and FLOB (from Freiburg University, Germany) contain American and British data from 1991; and ICE (International Corpora of English) contains regional varieties of English, such as Indian and Australian. Similar design principles underlie the Lund corpus of spoken British English (from University College London and Lund University), which contains around half a million words, divided into samples of the usage of adult, educated, professional people, including face-to-face and telephone conversations, lectures and discussions.

By the late 1990s, some corpora consisted of hundreds of millions of words. The Bank of English (at Cobuild in Birmingham, UK) and the British National Corpus

(BNC) had commercial backing from publishers, who have used the corpora to produce dictionaries and grammars. The 100-million word BNC is also carefully designed to include demographically and stylistically defined samples of written and spoken language. The Bank of English arguably over-emphasizes mass media texts, but these are very influential, and it still has a range of text-types and advantages of size: over 400-million words by 2001. Because constructing large reference corpora is so expensive, it may be that huge new corpora cannot again be created in the near future. These corpora will remain standard reference points, which can be supplemented by small specialized corpora, designed by individual researchers, and by large opportunistic collections.

Many other corpora for English, and increasingly for other languages, are available (see Michael Barlow's web-site: address below).

CORPUS DESIGN

Some basic principles of corpus design (Kennedy 1998, pp.13-87, Hunston 2002, pp.25-37) are simple enough. A corpus which claims to be a balanced sample of language use must represent variables of demography, style and topic, and must include texts which are spoken and written, casual and formal, fiction and non-fiction, which vary in level (e.g. popular and technical), age of audience (e.g. children or adults), and sex and geographical origin of author, and which illustrate a wide range of subject fields (e.g. natural and social sciences, commerce, and leisure). However, no corpus can truly represent a whole language, since no-one quite knows what should be represented. It is not even obvious what are appropriate proportions of mainstream text-types such as quality newspapers, literary classics and everyday conversation, much less text-types such as newspaper ads, business correspondence and church sermons. (Even carefully designed corpora have odd gaps: despite their influence as a text-type, textbooks are not represented in Brown and LOB.) A realistic aim is a corpus which samples widely, is not biased towards data which are easy to collect (e.g. mass media texts), does not under-represent data which are difficult to collect (e.g. casual conversation), and is not unbalanced by text-types which have over-specialized lexis and grammar (e.g. academic research articles).

Since large quantities of data are necessary in order to study what is typical and probable, an important criterion is size, which is usually measured in terms of running words (tokens). But measures of heterogeneity are also important: How large is the corpus measured as word-types (i.e. different words), or as the number of different texts or text-types it contains? A corpus might be very large, but consist entirely of American newswire texts, with a correspondingly narrow vocabulary. One can also attempt to measure linguistic influence: How large is the audience for the texts in the corpus? Casual conversation is a linguistic universal, but a typical conversation is private, whereas the language of the mass media is public, and therefore much more influential. And whereas some texts are heard once by millions of people (sports commentaries), others (literary classics) are

constantly re-read over generations. A reception index, which weights texts by their audience size, can be constructed at least in a rough way.

In summary, any corpus is a compromise between the desirable and the feasible, and although design criteria cannot be operationalized, large balanced corpora reveal major regularities in language use. In any case, there is no reason to rely on any single corpus, and it is often advisable to combine large general corpora designed according to principles of sociolinguistic variation, small corpora from specific knowledge domains (since much lexis is determined by topic), and opportunistic text collections.

Huge text collections (such as the world-wide-web) can be used to study patterns which do not occur even in large reference corpora. For example, concordance lines in the Appendix show that *undergo* is typically used of someone who is forced to undergo something unpleasant, often a medical procedure or a test of some kind, or of a situation which undergoes some profound and often unwelcome change. Typical examples are:

- had to *undergo* a stringent medical examination
- is about to *undergo* dramatic changes

However generalizations must be checked against potential counter-examples. First, comparison of different text-types shows that, in scientific and technical English, *undergo* usually has no unpleasant connotations. An example from the BNC (which still involves "change") is:

- the larvae *undergo* a complex cycle of 12 stages

Second, people "unwillingly" undergo unpleasant experiences. But does the collocation *willingly UNDERGO* occur and does it provide a counter-example? Now we have a problem: the lemma UNDERGO is fairly frequent (around 25 occurrences per million words in the BNC), and even *willingly* is not infrequent (around 5 per million), but the combination *willingly UNDERGO* does not occur at all in the 100-million word BNC. However, a search of the world-wide-web quickly provided 200 examples, which revealed another pattern: people *willingly undergo* a sacrifice for the sake of others or for the sake of religious beliefs. Characteristic examples are:

- one can *willingly undergo* some painful experience for one who is dearly loved
- sufferings and dangers the early Christians *willingly underwent* for the sake of ...

A corpus is specifically designed for language study, but other text collections (such as newspapers on CD-ROM) can be useful for some types of study. Again, I see no point in being purist about data, as long as their source is stated in a way

which allows findings to be assessed. The world-wide-web has the advantage of enormous size, but it is impossible to characterize its overall range of texts. Words and phrases in the world-wide-web can be searched for directly with search engines, or with a concordancer which uses these engines, such as one developed at the University of Liverpool (http://www.webcorp.org.uk/).

RAW, LEMMATIZED AND ANNOTATED CORPORA

A corpus may consist of raw text (strings of orthographic word-forms), or it can be lemmatized, and annotated or tagged, for intonation (for spoken corpora), grammatical or semantic categories. Part-of-speech tagging allows a corpus to be searched for grammatical constructions, such as adjective-noun combinations (*persistent rain*), and make it possible to study the frequency of grammatical categories in different text-types (e.g. see Biber et al 1998: pp.59-65 on nominalizations; and Carter and McCarthy 1999 on passives). Information on the frequencies of lexical and grammatical features can indicate to language teachers where it is worth while devoting pedagogical effort (Kennedy 1998, pp.88-203).

Nevertheless, a simple example illustrates the value of working with raw text. Many occurrences of the lemmas of the verbs PERSIST and ENDURE share the semantic and pragmatic features that something "unpleasant" is lasting "for a long time". However, although the adjectives *persistent* and *enduring* also share the feature "for a long time", their typical collocates show their very different connotations:

- persistent <ambiguity, bleeding, confusion, headaches>
- enduring <appeal, legacies, peace, significance, values>

Traditionally, lemmas comprise words within a single part of speech. *Persistent* is an adjective, and shares the connotations of the verb PERSIST. *Enduring* might be considered an adjective, or the *–ing* form of the verb ENDURE, but has very different connotations from the verb.

In addition, the grammatical categories needed for unrestricted naturally occurring text can be very different from those required for the invented data described in abstract syntax. This draws into question centuries-old assumptions about the part-of-speech system (Sinclair 1991, pp.81-98, Sampson 1995, Hallan 2001). So, tagging may make unwarranted assumptions about appropriate grammatical categories. Again, the principle is that observer and data should be kept independent. The facts never 'speak for themselves', but inductive methods aim for the minimum of preconceptions. How to lemmatize words is by no means always obvious, and there are no standardized systems for part-of-speech tagging (Atwell et al 2000) or full parsing (Sampson 1995).

4. NEW FINDINGS AND DESCRIPTIONS

The main findings which have resulted from the 'vastly expanded empirical base' (Kennedy 1998, p.204) which corpora provide concern the association patterns which inseparably relate item and context:

- lexico-grammatical units: what frequently (or never) co-occurs within a span of a few words
- style and register: what frequently (or never) co-occurs in texts.

Findings about lexico-grammar question many traditional assumptions about the lexis-grammar boundary. The implications for language teaching are, at one level, rather evident. A well known problem for even advanced language learners is that they may speak grammatically, yet not sound native-like, because their language use deviates from native-speaker collocational norms. I once received an acknowledgement in an article by a non-native English-speaking colleague, for my 'repeated comments on drafts of this paper', which seemed to connote both irritation at my comments and to imply that they were never heeded. (I suppose this was better than being credited with 'persistent comments'!)

Syllabus designers ought to know which words are used frequently in conventionalized combinations, and which are used rarely and in special contexts. The importance of collocations for language learners was emphasized in the 1930s and 1940s by H. E. Palmer and Hornby. More recently corpora have been used to study how learners and native speakers differ in their use of conventionalized expressions (Granger ed 1998), and a major topic has been how to represent such information in learners' dictionaries (Cowie ed 1998). Proposals have also been made about the form of a 'lexical syllabus'. This concept was discussed in detail by Corder (1973, pp.315-17), and has been revived in corpus work by Willis (1990) and Lewis (1998), although corresponding teaching materials have been adopted only to a limited extent. The shorthand label for this area is phraseology: the identification of typical multi-word units of language use and meaning.

WORDS

Many corpus studies reject individual words as units of meaning, and propose a theory of abstract phrasal units. Nevertheless, words are a good place to start, since, 'a central fact about a word is how frequent it is' (Kilgarriff 1997, p.135), and other things being equal, the more frequent a word is, the more important it is to know it, and to teach it early to learners: hence the interest, since the 1890s, in reliable word-frequency lists for many applications.

Frequency shows that system and use are inseparable (Halliday 1991). More frequent words tend to be shorter, irregular in morphology and spelling, and more ambiguous out of context: a glance at a dictionary shows that short frequent words

require many column inches. A few, mainly grammatical, words are very frequent, but most words are very rare, and an individual text or smallish corpus, around half the words typically occur only once each. In addition, a word with different senses usually has one meaning which is much more frequent. These relations imply a balance between economy of effort for the speaker and clarity for the hearer, and in the 1930s and 1940s Zipf (1945) tried to formulate statistical relations between word frequency, word length and number of senses. (These regularities apply to many other aspects of human behaviour. In a library, a few books are frequently borrowed, but most books collect dust.)

The simplest frequency lists contain unlemmatized word-forms from a general corpus, in alphabetical or frequency order, but there are considerable differences between even the top ten words from an unlemmatized written corpus (in 1), a spoken corpus (in 2), and a lemmatized mixed written and spoken corpus (in 3):

(1) the, of, and, a, in, to [infinitive marker], is, to [preposition], was, it
(2) I, you, it, the, 's, and, n't, a, that, yeah
(3) the, BE, of, and, a, in, to [infinitive marker], HAVE, it

These examples are from frequency lists for the 100-million word BNC, made available by Kilgarriff (ftp://ftp.itri.bton.ac.uk/bnc/).

Unlemmatized lists show that different forms of a lemma differ greatly in frequency, and may have very different collocational behaviour: see above on *endure* and *enduring*. However, raw frequency lists cannot distinguish words in different grammatical classes (e.g. *firm* as adjective or noun) and the different meanings of a word (e.g. *cold* as "low temperature" versus "lacking in feeling"). This requires a grammatically tagged corpus and a method of automatic sense disambiguation, and makes an apparently trivial counting task into a considerable theoretical problem.

Frequency lists require careful interpretation to provide what is really wanted, which is a measure of the relative importance of words, and more important than raw frequency may be even distribution across many text-types. Conversely, we want to know not only what is frequent in general, but what distinguishes a text-type. For example, words may be frequent in academic texts but unlikely in fiction, or vice-versa:

- constants, measured, thermal, theoretically
- sofa, kissed, damned, impatiently

These examples are from Johansson (1981, discussed also by Kennedy 1998, p.106). For important reference data on word-frequency and distribution, see W. Francis and Kucera (1982), Johannson and Hofland (1988-89), and Leech et al (2001, and http://ucrel.lancs.ac.uk/bncfreq/ [accessed Jan 2016]).

We come back to the distinction between evidence and interpretation. Frequency and distribution (which are all we have) are indirect objective measures of the subjective concept of salience (which is what we really want). The objective measures have limitations, but allow analysis to be based on public and replicable data. The only alternative is intuition, which may be absent, speculative or wrong.

A very useful applied frequency study is reported by Coxhead (2000), who used a corpus of 3.5 million words to set up the Academic Word List (AWL). This contains words which have both high frequency and wide distribution in academic texts, irrespective of subject area (but excluding approximately the 2,000 most frequent words in English, from West 1953). AWL comprises 570 word families: not just word-forms, but head-words plus their inflected and derived forms (see above), and therefore around 3,100 word-forms altogether, e.g.:

- concept: conception, concepts, conceptual, conceptualisation, conceptualise, conceptualised, conceptualises, conceptualising, conceptually.

Coxhead's corpus comprised texts from academic journals and university textbooks from arts, commerce, law, and natural science. To be included in AWL, a word had to occur at least 100 times altogether in the whole academic corpus, at least ten times in each of the four sub-corpora, and in at least half of 28 more finely defined subject areas, such as biology, economics, history, and linguistics. AWL gives very good coverage of academic texts, irrespective of subject area. Here it must be remembered that words are *very* uneven in their frequency. In a typical academic text, the single word *the* covers around 6 or 7 per cent of running text, the top ten words cover over 20 per cent, and the 2,000 most frequent words cover around 75 per cent. The words in AWL typically cover a further 10 per cent. The remaining 15 per cent will be specialized words which are specific to a given topic, plus proper names, etc. AWL is further divided into ten sub-groups, from most to least frequent. Group 1 covers 3.6 per cent of the corpus, which means that a student is reading academic prose could expect to come across *each word* in group 1, on average, once every four pages or so.

A list is, of course, just a list, not teaching materials, and requires interpretation by materials designers and teachers. However, even as a bare list, AWL can provide a check, for teachers or students themselves, on what words students should know.

PHRASES

Word frequency lists are limited, especially for very common words, since these are common, not in their own right, but because they occur in common phrases. For example, *back* is usually in the top 100 in lemmatized frequency lists, and (including compounds such as *backward* and *backwater*) gets nearly five full pages in the Cobuild (1995a) dictionary. This is not because speakers frequently use *back* to mean a part of the body, but because it occurs in many phrases with only residual relations to this denotation. It has many meanings, but vanishingly

few uses with the part-of-body meaning. The following examples are from Cobuild (1995a), and Sinclair (1991: 116) gives a detailed analysis of its nominal, prepositional and idiomatic uses.

- lying on his back; the back of the chair; on the back of a postcard; at the back of the house; round the back; do something behind her back; get off my back; you scratch my back ...; see the back of someone; turn your back on

In summary: Frequent words are frequent because they occur in frequent phrases. In these phrases, frequent words are often delexicalized, because meaning is dispersed across the whole phrase. Since frequent content words are rarely used with their full lexical meaning, the boundary between content and function words is fuzzy. It is for these reasons that the co-occurrence of words and grammatical constructions has been studied so intensively: the central principle is that it is not words, but phrase-like units, which are the basic units of meaning.

RECURRENT PHRASES, COLLOCATIONS AND PHRASAL SCHEMAS

The simplest definition of a phrase is a string of two or more uninterrupted word-forms which occur more than once in a text or corpus: see Altenberg (1998) on 'recurrent word-combinations' and Biber et al (1999) on 'lexical bundles'. I used a program to identify strings in this sense, in a written corpus of four million words. The most frequent five-word string, over twice as frequent as any other, was *at the end of the*. And almost 30 out of the top 100 five-word strings had the pattern *PREP + the + NOUN + of + the*. Examples included:

- at the end of the; in the middle of the; at the beginning of the; at the bottom of the

The program operationalizes, in a very simple way, the concept of repeated units. It cannot automatically identify linguistic units, but presents data in a way which helps the analyst to see patterns. These findings are not an artefact of my small corpus. I looked at the same strings in the 100-million word BNC, and found that, normalized to estimated occurrences per million words, the frequencies in the two corpora were remarkably similar. These examples represent only one pattern, of course. Other frequent five-word strings have discourse functions:

- as a matter of fact; it seems to me that; it may well be that; but on the other hand

Altenberg (1998) identifies other recurrent multi-word strings, and some of their typical pragmatic functions.

These multi-word strings are already evidence that recurrent lexico-grammatical units are not fixed phrases, but abstract semantic units. For example, the program

above counts separately the strings *on the top of the*, *on the very top of the* or *on top of the*, although, to the human analyst, they are semantically related.

More abstract again is the concept of collocation, in the sense of the habitual co-occurrence of word-forms or lemmas. A few dozen concordance lines can be manually inspected for patterns, but if we have thousands of lines, then we require a method of summarizing concordances and showing patterns. We can write a program which finds the most frequent collocates of a node, one, two and three words to the left and right, and lists them in descending frequency. The positional frequency table for *undergo* shows that it often occurs in a passive construction (*was forced to*, *is required to*), is often followed by an adjective signalling the seriousness of the event (*extensive*, *major*), and is often used of medical events (*surgery*, *operation*).

INSERT POSITIONAL FREQUENCY TABLE for undergo ABOUT HERE

Raw frequency of co-occurrence is important, but we need to check the frequency of collocation relative to the frequency of the individual words. If two words are themselves very frequent, they may co-occur frequently just by chance. Conversely, a word might be infrequent, but when it does occur, it usually occurs with a small set of words. For example, the word *vegetative* is not frequent, but when it occurs, especially in journalism, it often co-occurs with *persistent*, in the phrase *persistent vegetative state*, with reference to patients in a coma.

The variability of phrasal units makes it doubtful whether there could be a useful 'phrase frequency list', but corpus studies show that all words occur in habitual patterns which are often much stronger than is evident to intuition. For example, in a 200-million word corpus, the word-form *persistent* occurred over 2,300 times, with clear semantic preferences, shown by the top 20 collocates, ordered by frequency:

- persistent <offenders, reports, most, rumours, state, vegetative, despite, young, juvenile, problem, injury, problems, rain, allegations, critic, offender, rumors, speculation, amid, cough>

The most frequent single collocate (in 5 per cent of cases) was *offenders*; and the most frequent set of collocates were words for *reports*, *rumours* and *speculations*. *Persistent* is used of bad situations (collocates include *problem* and *problems*), which include medical conditions (*cough*, *injury*, *vegetative*) and criminal activities (*juvenile*, *offenders*). Some collocates frequently occur in longer phrases (*persistent juvenile offenders*, *persistent vegetative state*), and most examples involving "crime" and "allegations" are from journalism. With comparable data on a broad sample of words, we can then ask whether *persistent* exerts a stronger

than average collocational attraction on its surrounding collocates. The brief answer is that *persistent* is typical of many words in this respect.

The top collocates of a word provide evidence of its characteristic semantic preferences and syntactic frames. Figures for a broad sample of words show how pervasive collocational attraction is, and allow generalizations about its strength and variability. The example of *persistent* is taken from a data-base (Cobuild 1995b), which provides a suitable sample of node-words and their collocates for quantitative statements about phraseology. For the 10,000 most frequent content words (word-forms) in the 200-million word corpus, the data-base gives the 20 most frequent collocates in a span of four words to left and right. For each node-collocate pair, it gives 20 randomly selected concordance lines, each with a rough description of its source (e.g. British fiction, American journalism). For individual words, this provides figures on the strength of attraction between node and top collocate:

- undergoing <surgery 11%>, undergo <surgery 9%>, endured <years 6%>, persistent <offenders 5%>

(That is, in 11 per cent of occurrences, *undergoing* co-occurs with *surgery*, etc.) The data-base shows that around 75 per cent of content words in the central vocabulary of English have a strength of attraction of between 2 and 9 per cent. And over 20 per cent co-occur with one specific collocate in over 10 per cent of occurrences. Conversely, few words have less than one chance in 50 of co-occurring with one specific collocate.

These are figures for the attraction between two single unlemmatized word-forms. Collocational attraction is much stronger if it is calculated between a node and a set of approximate synonyms. For example:

- achieving <goal(s) 7%, success, aim, results, objectives> 15%
- ambitious <plan(s) 7%, project, program(me), scheme> 16%

The strength of attraction between all common content words is surprisingly high, yet not taken into account in most language description. Corpus study shows kinds of linguistic organization which are not predictable by rule, but are recurrent and observable.

SEMANTIC PREFERENCE, DISCOURSE PROSODY AND EXTENDED LEXICAL UNITS

A central aim is to make more explicit the semantic and pragmatic features of multi-word units. For example, *enduring*, *persistent* and *haunting* are all rough synonyms, which share a propositional meaning, but they co-occur with nouns from different semantic fields and have different evaluative connotations. Characteristic combinations of modifier plus noun include:

- enduring peace; haunting music; persistent headaches

We can also generalize about semantic preferences. In adjective-noun constructions, *persistent* is often used of medical conditions, and *haunting* is usually used of music, words and images. Different speaker attitudes are also conveyed: *persistent* is used of unpleasant topics, whereas *enduring* and *haunting* are usually used of things which are valued. For some speakers, ENDURE will have further Biblical connotations, since it occurs frequently in the King James translation: often with positive connotations when intransitive (*his mercy endureth for ever*), and often negative transitive (*endureth temptation*). Louw (1993) was the first important article on how such attitudes are conveyed.

A model of extended lexical units proposed by Sinclair (1998) combines these increasingly abstract relations: (1) collocation (the habitual co-occurrence of individual word-forms or lemmas), (2) colligation (the co-occurrence of words and grammatical categories), (3) semantic preference (the co-occurrence of a word or grammatical construction with words from a well defined semantic field), and (4) discourse prosody (a descriptor of speaker attitude and discourse function). We can also specify: (5) strength of attraction between node and collocates; (6) position of node and collocate, variable or fixed (as in *spick and span*, but not **span and spick*); and (7) distribution, wide occurrence in general English or in broad varieties (e.g. journalism), or restricted to specialized text-types (e.g. recipes: *finely chopped*; or weather forecasts: *warm front*).

In summary: Work on extended lexical units has redrawn the lexis-grammar boundary. Only a few units are fixed phrases; most are recurrent combinations of grammatical constructions with words from restricted lexical fields, but with considerable lexical variation. A good term is 'stabilized expressions' (Lenk 2000). So, the vocabulary of a language is not merely 'a list of basic irregularities' (Bloomfield 1933, p.274). Relations (1) to (4) correspond to the classic distinctions between syntax (how language units relate to one another), semantics (how linguistic signs relate to the external world), and pragmatics (how linguistic signs relate to their users, here expression of speaker attitude). This model has profoundly influenced dictionary design (Cowie ed 1998) and language teaching (Hunston 2002).

GRAMMAR, CO-TEXT AND TEXT-TYPE

Corpus work has taken the development of grammars in two directions: description of the pervasive co-selection of grammar and lexis, and of grammatical variation in different text-types.

The examples above of lexico-grammatical units illustrate very briefly the type of patterns which G. Francis et al (1996, 1998) document systematically in the first corpus-driven grammars of English. For each verb, noun and adjective in a large corpus, down to a frequency cut-off point, they show 'the patterns that are

associated with particular lexical items' (Hunston & Francis 2000, p.1). These highly innovative grammars show, for the first time, across the whole language, the intimate interaction between lexis, grammar and meaning. Starting from individual words, users can find the grammatical patterns in which the words typically occur. Starting from the grammar, users can find the semantically related words which typically occur in the patterns, and therefore the meanings which they typically express.

Corpus methods can also reveal characteristics of whole texts and text-types, such as what proportion of a text consists of repetitions of the same words or new words (its type-token ratio), the ratio of content to function words (its lexical density), or the relative proportions of everyday and academic vocabulary, and can establish the central tendencies and range of variation across text-types. Other things being equal, high type-token ratio, high lexical density, and high percentages of academic vocabulary will make a text more difficult to understand. Biber (1988) used quantitative and distributional techniques to identify words and grammatical constructions which frequently (or never) co-occur in text-types such as conversation, personal letters, and science fiction, and to identify textual dimensions such as informational, narrative, and persuasive.

The grammar of spoken and written English by Biber et al (1999), based on a 40-million word corpus of British and American English, shows the frequency and distribution of lexical and grammatical structures in different text-types. Taking just one specific finding, of great potential interest to anyone concerned with designing English language teaching materials, the grammar identifies (pp.373ff) the twelve most frequent lexical verbs in English. These are activity verbs (*get*, *go*, *make*, *come*, *take*, *give*), mental verbs (*know*, *think*, *see*, *want*, *mean*) and a communication verb (*say*). As a group, these verbs make up only 11 per cent of lexical verbs in academic prose, but nearly 45 per cent in conversation. Such findings do not translate directly into teaching materials or lesson plans, and applications of such work are still relatively modest, but such grammars indicate aspects of language use on which teachers may need to concentrate.

Although description of language use is inevitably description of language variation, G. Francis et al (1996, 1998) do not distinguish text-types, and Biber et al (1999) differentiate only four broad categories (conversation, fiction, newspaper language, academic prose). Given their need to present 'general English', dictionaries and grammars can take only limited account of variation within the language, and, as noted above, it is doubtful whether varieties can be exhaustively classified.

5. APPLICATIONS, IMPLICATIONS AND OPEN QUESTIONS

There are often striking differences between earlier accounts of English usage (pedagogical and theoretical) and corpus evidence, but the applications of corpus findings are disputed. Since I cannot assess the wide range of proposed, rapidly changing and potential applications, I have tried to set out the principles of data

design and methods which applied linguists can use in assessing descriptions and applications. Perhaps especially in language teaching, one also has to assess the vested interests involved: both resistance to change by those who are committed to ways of teaching, and also claims made by publishers with commercial interests in dictionaries and teaching materials.

Apart from language teaching and lexicography, other areas where assessment is required are as follows:

(1) Translation studies. By the late 1990s, bilingual corpora and bilingual corpus-based dictionaries had developed rapidly. The main finding (Baker 1995, Kenny 2001) is that, compared with source texts, the language of target texts tends to be 'simpler', as measured by lower type-token ratios and lexical density, and the proportion of more explicit and grammatically conventional constructions.

(2) Stylistics. Corpora are the only objective source of information about the relation between instance and norm, and provide a concrete interpretation of the concept of inter-textuality. Burrows (1987) is a detailed literary case study, and Hockey (2001) discusses wider topics. The next category might be regarded as a specialized application of stylistics.

(3) Forensic linguistics. Corpus studies can establish linguistic norms which are not under conscious control. Although findings are usually probabilistic, and an entirely reliable 'linguistic fingerprint' is currently unlikely, corpus data can help to identify authors of blackmail letters, and test the authenticity of police transcripts of spoken evidence. Progress has also been made with other kinds of text comparison, such as identifying plagiarism and copyright violation. (Coulthard 1994.)

(4) Cultural representation and keywords. Several studies investigate the linguistic representation of culturally important topics: see Gerbig (1997) on texts about the environment, and Stubbs (1996) and Piper (2000) on culturally important keywords and phrases. Atkinson (1999) combines computational, manual and historical methods in a detailed study of an influential corpus of scientific writing from the 17th to the 20th century. Channell (2000) shows the importance of correctly representing the cultural connotations of cultural keywords in learner dictionaries.

(5) Psycholinguistics. On a broader interpretation of applications, psycholinguistic studies of fluency and comprehension can use findings about the balance of routine, convention and creativity in language use (Wray 2002). Corpus-based studies of child language acquisition have also questioned assumptions about word-categories and have far-reaching implications for linguistic description in general (Hallan 2001).

(6) Theoretical linguistics. The implications here lie in revisions or rejection of the *langue*/*parole* opposition, the demonstration that the tagging and parsing of

unrestricted text requires changing many assumptions about the part-of-speech system (Sinclair 1991, pp.81-98, Sampson 1995), and about the lexis/grammar boundary (G. Francis et al 1996, 1998).

Computer-readable corpora became available only in the 1970s, and for many years were limited and inconvenient. They became widely accessible only from the mid-1990s, when linguistics suddenly went from a position of being 'starved of adequate data' (Sinclair 1991, p.1) to being swamped with data. Development is now (post-2000) very rapid, but it will take time before we can see the wood for the trees, and state with certainty the long-term implications. No linguists can now ignore corpus data. Many severe difficulties in observing language use have been resolved, and although language corpora are not the only way of seeing language, they are a very productive way. With reference to language description, I have taken an enthusiastic view, arguing that language corpora have provided many new findings about lexis, grammar and semantics. With reference to applications, I have taken a conservative view, arguing that applications are indirect, and that, before findings can be applied to real-world problems, they require careful interpretation.

ACKNOWLEDGEMENTS

For pointing out to me the interest of the set of words related to PERSIST and PERSEVERE, I am grateful to Alan Partington (1996, pp.77, 80) and to my student Anne Schmidt. For writing the 'strings' and positional frequency programs, I am grateful to my student research assistants, Isabel Barth and Oliver Hardt.

RESOURCES AND FURTHER READING

1. Web-sites: corpus linguistics and corpora

These individual URLs were up-dated in Jan 2016. But the whole list is rather out-of date, since the article was published in 2004.

Corpus linguistics web-site (Michael Barlow):
http://michaelbarlow.com/.
Corpus linguistics web-site (David Lee):
http://tiny.cc/corpora.
Data-driven learning page (Tim Johns):
[no longer available?]
BNC (British National Corpus):
http://www.natcorp.ox.ac.uk/.
COBUILD (Collins Birmingham University International Language Database):
[no longer available?]
ICAME (International Computer Archive of Modern and Medieval English):
http://clu.uni.no/icame/.

LDC (Linguistic Data Consortium):
https://www.ldc.upenn.edu/
ICE (International Corpus of English):
http://ice-corpora.net/ice/index.htm.
Oxford Text Archive:
http://ota.ox.ac.uk/

2. Journals

Computers and the Humanities (1960s-)
ICAME Journal (1976-, previously *ICAME News*)
International Journal of Corpus Lingustics (1996-)
Literary and Linguistic Computing (1986-, in its present form)

3. Textbooks

Barnbrook, G. (1996). *Language and computers: a practical introduction to the computer analysis of language*. Edinburgh: Edinburgh University Press.
Partington, A. (1998). *Patterns and meanings: using corpora for English language research and teaching*. Amsterdam: Benjamins.
Stubbs, M. (2001). *Words and phrases: corpus studies of lexical semantics*. Oxford: Blackwell.

4. Articles

Barlow, M. (1996). Corpora for theory and practice. *International Journal of Corpus Linguistics*, 1, 1: 1-37.
Biber, D., Conrad, S. & Reppen, R. (1994). Corpus-based approaches to issues in applied linguistics. *Applied Linguistics*, 15, 2: 169-89.
Cowie, A. P.(1999). Phraseology and corpora: some implications for dictionary-making. *International Journal of Lexicography*, 12, 4: 307-23.
De Beaugrande, R. (2000). Text linguistics at the millenium: corpus data and missing links. *Text*, 20, 2: 153-95.
Fillmore, C. J. & Atkins, B. T. S. (1994). Starting where the dictionaries stop: the challenge of corpus lexicography. In B. T. S. Atkins & A. Zampoli (Eds.) *Computational approaches to the lexicon*. (pp.349-93). Oxford: Clarendon.
Pawley, A. (2001). Phraseology, linguistics and the dictionary. *International Journal of Lexicography*, 14, 2: 122-34.

REFERENCES

Altenberg, B. (1998). On the phraseology of spoken English. In A. P. Cowie (Ed.) *Phraseology: theory, analysis and applications*. (pp.101-122). Oxford: Oxford University Press.
Atkinson, D. (1999). *Scientific discourse in sociohistorical context*. Mahwah, NJ: Erlbaum.

Atwell, E., Demetriou, G., Hughes, J., Schiffrin, A., Souter, C. & Wilcock, S. (2000). A comparative evaluation of modern English corpus grammatical annotation schemes. *ICAME Journal*, 24: 7-24.

Ayscough, S. (1790). *An index to the remarkable passages and words made use of by Shakespeare*. London: Stockdale.

Baker, M. (1995). Corpora in translation studies. *Target*, 7, 2: 223-43.

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

Biber, D., Conrad, S. & Reppen, R. (1998). *Corpus linguistics.* Cambridge: Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. London: Longman.

Bloomfield, L. (1933). *Language*. London: Allen & Unwin, 1935.

Burrows, J. F. (1987). *Computation into criticism*. Oxford: Clarendon.

Carter, R. & McCarthy, M. (1999). The English *get*-passive in spoken discourse. *English Language and Linguistics*, 3, 1: 41-58.

Channell, J. (2000). Corpus-based analysis of evaluative lexis. In S. Hunston & G. Thompson (Eds.) *Evaluation in text*. (pp.38-55). Oxford: Oxford University Press.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, Ma.: MIT Press.

Chomsky, N. (2000). *New horizons in the study of language and mind*. Cambridge: Cambridge University Press.

Church, K. W. & Hanks, P. (1990). Word association norms, mutual information and lexicography. *Computational Linguistics*, 16, 1: 22-29.

CIDE (1995). *Cambridge international dictionary of English*. Cambridge: Cambridge University Press.

Clear, J. (1993). From Firth principles: computational tools for the study of collocation. In M. Baker, G. Francis & E. Tognini-Bonelli (Eds.) *Text and technology*. (pp.271-92). Amsterdam: Benjamins.

COBUILD (1990). *Collins COBUILD English grammar*. London: HarperCollins.

COBUILD (1995a). *Collins COBUILD English dictionary*. London: HarperCollins.

COBUILD (1995b). *Collins COBUILD English collocations on CD-ROM*. London: HarperCollins.

Corder, P. (1973). *Introducing applied linguistics*. Harmondsworth: Penguin.

Coulthard, M. (1994). On the use of corpora in the analysis of forensic texts. *Forensic Linguistics*, 1: 27-44.

Cowie, A. P. (Ed.) (1998). *Phraseology: theory, analysis and applications*. Oxford: Oxford University Press.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, 34, 2. 213-38.

Cruden, A. (1737). *A complete concordance to the Holy Scriptures*. 10th edition, 1833. London: Tegg.

De Beaugrande, R. (1999). Reconnecting real language with real texts: text linguistics and corpus linguistics. *International Journal of Corpus Linguistics*, 4, 2: 243-59.

Fillmore, C. J. (1992). Corpus linguistics or computer-aided armchair linguistics. In J. Svartvik (Ed.) *Directions in corpus linguistics*. (pp.35-60). Berlin: Mouton.

Francis, G., Hunston, S. & Manning, E. (1996, 1998). *Grammar patterns*. 2 volumes: *Verbs*, and *Nouns and adjectives*. London: HarperCollins.

Francis, W. N. (1992). Language corpora BC. In J. Svartvik (Ed.) *Directions in corpus linguistics*. (pp.17-32). Berlin: Mouton.

Francis, W. N. & Kucera, H. (1982). *Frequency analysis of English usage: lexicon and grammar*. Boston: Houghton Mifflin.

Fries, C. C. (1952). *The structure of English*. NY: Harcourt, Brace & World.

Gerbig, A. (1997). *Lexical and grammatical variation in a corpus*. Frankfurt: Lang.

Granger, S. (Ed.) (1998). *Learner English on computer*. London: Longman.

Hallan, N. (2001). Paths to prepositions? A corpus-based study of the acquisition of a lexico-grammatical category. In J. Bybee & P. Hopper (Eds.) *Frequency and the emergence of linguistic structure*. (pp.91-120.) Amsterdam: Benjamins..

Halliday, M. A. K. (1991). Corpus studies and probabilistic grammar. In K. Aijmer & B. Altenberg (Eds.) *English corpus linguistics*. (pp.30-43). London: Longman.

Hockey, S. (2001). *Electronic texts in the humanities*. Oxford: Oxford University Press.

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

Hunston, S. & Francis, G. (2000). *Pattern grammar: a corpus-driven approach to the lexical grammar of English*. Amsterdam: Benjamins.

Jespersen, O. (1909-49). *Modern English grammar*. Heidelberg: Winter. Vols 1-4. Copenhagen: Munksgaard. Vols 5-7.

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

Johansson, S. & Hofland, K. (1988-89). *Frequency analysis of English vocabulary and grammar*. 2 volumes. Oxford: Oxford University Press.

Kaeding, F. W. (1898). *Häufigkeitswörterbuch der deutschen Sprache*. Berlin: Steglitz.

Kennedy, G. (1998). *An introduction to corpus linguistics*. London: Longman.

Kenny, D. (2001). *Lexis and creativity in translation: a corpus-based study*. Manchester: St Jerome.

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

Leech, G., Rayson, P. & Wilson, A. (2001). *Word frequencies in written and spoken English*. London: Longman.

Lenk, U. (2000). Stabilized expressions in spoken discourse. In C. Mair & M. Hunt (Eds.) *Corpus linguistics and linguistic theory*. (pp.187-200). Amsterdam: Rodopi.

Lewis, M. (1998). *Implementing the lexical approach*. Hove: Language Teaching Publishers.

LDOCE (1995). *Longman dictionary of contemporary English*. London: Longman.

Louw, B. (1993). Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In M. Baker, G. Francis & E. Tognini-Bonelli (Eds.) *Text and Technology*. (pp.157-76). Amsterdam: Benjamins.

OALD (1995). *Oxford advanced learner's dictionary*. Oxford: Oxford University Press.

Piper, A. (2000). Lifelong learning, human capital and the soundbite. *Text*, 20, 1: 109-46.

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. (1972). *A grammar of contemporary English*. London: Longman.

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. (1985). *A comprehensive grammar of the English language*. London: Longman.

Sampson, G. (1995). *English for the computer*. Oxford: Clarendon.

Sinclair, J. (1991). *Corpus concordance collocation*. Oxford: Oxford University Press.

Sinclair, J. (1998). The lexical item. In E. Weigand (Ed.) *Contrastive lexical semantics*. (pp.1-24). Amsterdam: Benjamins.

Sinclair, J., Mason, O., Ball, J. & Barnbrook, G. (1998). Language independent statistical software for corpus exploration. *Computers and the Humanities*, 31: 229-55.

Stubbs, M. (1996). *Text and corpus analysis*. Oxford: Blackwell.

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

Thorndike, E. L. & Lorge, I. (1944). *A teacher's word book of 30,000 words*. NY: Teachers' College, Columbia University.

West, M. (1953). *A general service list of English words*. London: Longman.

Willis, D. (1990). *The lexical syllabus*. London: Collins.

Wilson, E. O. (1998). *Consilience: the unity of knowledge*. London: Little, Brown & Co.

Winchester, S. (1998). *The surgeon of Crowthorne: a tale of murder, madness and the Oxford English Dictionary*. London: Viking.

Wray, A. (2002). *Formulaic language and the lexicon*. Stanford, CA: Cambridge University Press.

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

POSITIONAL FREQUENCY TABLE for NODE *undergo* in a span of 3 words to left and right. Only collocates occurring 5 or more times are shown, in descending frequency, independently for each position.

N-3	N-2	N-1	NODE	N+1	N+2	N+3
was	forced	to	*	a	medical	and
is	required	will	*	an	surgery	tests
be	have	and	*	further	testing	examination
are	had	would	*	extensive	tests	of
and	is	must	*	the	treatment	surgery
that	they	he'll	*	major	change	operation
been	about	should	*	surgery	changes	transformation
were	and	who	*	treatment	for	before
where	patients	women	*	medical	heart	test
children	that	often	*	heart	and	medical
he	he		*	his	major	for
in	will		*	testing	operation	in
the	women		*		examination	on
women	due		*		extensive	training
will	ordered		*		transformation	to
for			*		radical	testing
last			*		test	the
not			*		training	a
of			*		the	as
			*			by
			*			changes

APPENDIX. ILLUSTRATIVE CONCORDANCE DATA.

These are a very few attested, but purely illustrative, concordance lines. They are not a random or representative sample of the corpora from which they are drawn. Readers could however study larger samples of the node words from other corpora and check whether they find comparable examples, and could also check whether other word-forms of the lemmas (e.g. *endures*, *endured*) show the same patterns. These examples are taken from the publically accessible versions of CobuildDirect and the BNC. The concordance lines are ordered alphabetically to the right of the node word.

Word-forms *endure*, *persevere*, *persist* and *undergo*.

```
01  st that smokers will have to endure 12-hour flights by becoming mo
02  d can remember having had to endure a certain amount of misery bef
03  ng that Romania still had to endure a period of austerity. Rome
04  ht find himself compelled to endure a spartan existence; unlike a
05  so that the rider has had to endure a steady worsening of the trav
06  erced family audience has to endure an hour of his old cine films,
07   the 1,700 prisoners have to endure constant noise from the Garmen
08  dertake forced labour and to endure dehumanizing captivity in the
09  t workers in El Paso, Texas, endure difficult conditions, and comp
10  e felt he had been forced to endure during the last three years. I
11  he birth. These episodes may endure for a few days or may linger f
12  nd the animals often have to endure hours trapped in the midst of
```

do nothing about, other than endure it or enjoy it, but it is alwa
lame. At last, when he could endure no more, he jerked his hands a
in a dark and cold place, to endure patiently sorrow and weakness
ans, for they were forced to endure the indignity of having anothe
over, one finds it easier to endure those tedious weekly audiences
aving to accept and bear and endure, and because I am quite clever
s will be painful for her to endure, and for you to witness, but u
ment. But they persevere and endure, rather than come out

ying at the moment. But they persevere and endure, rather than com
 to quit, half determined to persevere he was caught for some mome
mething about the ability to persevere in adversity. Koppel: Well,
t they produce. And we shall persevere in our efforts to find the
 them to concentrate on, and persevere in solving problems and pur
atient's family as a need to persevere in the face of inevitable l
ing and difficult but if you persevere in the most important area
raiseworthy, and urge you to persevere in this work of salvation.
ting Colonel North failed to persevere through adversity or anythi
 determined to remain and to persevere until she reaches a working
 is often quite difficult to persevere with tape-recording during
t completely. Be patient and persevere with the inoculation - it m
game to get into, but if you persevere you won't be disappointed.
ts were fully determined 'to persevere' with the three-strand form
 stage, but Brian decided to persevere, moving the boat to EDJ Boa
earliest efforts, but should persevere, using a single rock sample
 the ability to do it. If we persevere, we will get there. I accep
destroyed his willingness to persevere, yet since Izzy's reawakeni
who insisted that she should persevere. One was a bright editor at
 do this and it works if you persevere. You need to work at it - i

nsiderable misunderstandings persist about the nature of the handi
hat tremendous uncertainties persist about the relative importance
appropriate if the movements persist and are causing the child an
operation, and that this can persist for five years or more. For
ally cold temperatures might persist for over a year. Any survivor
n that residual activity may persist for up to six weeks following
is it that many commentators persist in calling the Presocratics s
portunity, should the regime persist in its ill-advised campaign a
the region, parents will not persist in the face of the child's re
r-pistol if the dog tries to persist in this antisocial behaviour.
ingle wet straws. Why do you persist in this perversity? Why do yo
d, dead batteries and if you persist in trying to recharge an
 the office governments will persist in trying to regulate what we
 minor ailments. If symptoms persist or are severe please consult
 like Julie Andrews. Rumours persist that her brother will join he
ny smooth passages but fears persist that modern lightweight racin
l three weeks ago. And fears persist that the PLO too may be drift
is not successful, he should persist until he has got what he want
mpassable forest, but if you persist you may find, depending on re
orth but the light rain will persist, especially over high ground.

lued women would have had to undergo a deep and important change o
he old people were likely to undergo a major psychological upheava
driving, had been induced to undergo a medical examination to see
 work, each operative had to undergo a stringent medical examinati
racter of the shop seemed to undergo a transformation. The rush wa
ate. Mr Forbes was forced to undergo an emergency operation to rem
dly take kindly to having to undergo an identity check before bein
tually anyone at risk should undergo confidential testing on a tra
hospital and insisted that I undergo extensive tests. There was he
officers and men have had to undergo great privations. They landed
cope with two recessions and undergo immense change in that proces
     Many of these creatures undergo intolerably cruel conditions

titute employees may have to undergo lie detector tests. Rapist w
fractured skull may have to undergo neuro-surgery if his conditio
g, if they were expecting to undergo surgery, or if they had a his
ho find themselves having to undergo the painful dislocation entai
ur means he will not have to undergo the punishing marathon of the
ronization, and initiative - undergo trial by fire. Holder also ha
but they would also need to undergo years of specialized training
ree RAF widows would have to undergo 'demeaning means tests' years

Word-forms *enduring, haunting* and *persistent* followed immediately by a noun.

andist only testified to his enduring ability to draw a crowd. 53
becoming a smash hit. The enduring appeal of Unchained Melody to
easoned optimism and by their enduring courage press on when lesser
also fails to reflect the enduring fascination of sporting it is
is the SUN which provides an enduring image of how Mrs Thatcher has
daily lives Perhaps the most enduring legacy of Thatcherism is that
ries, for the prestige or the enduring legacy of having their name o
goofing around, it's about an enduring love of guitars that borders
rary education with a work of enduring merit from Everyman's Library
intended to study music, an enduring passion of his which is refle
Hampshire's winsome charm and enduring popularity have elicited pity
the all-time bestsellers. Its enduring popularity is beyond doubt, a
of the credit for `Messiah's" enduring popularity belongs to the
rticular, Raeder developed an enduring reverence for the Baumeister
overworked person. Given the enduring sense of identity within
al" forms of masculinity, the enduring significance of the power of
of 1945 was led by men of enduring stature. Do you believe that
ars, this tree will become an enduring symbol of your commitment to
OUS Kelly Brown displayed her enduring talent when winning the Silk
in the 5th century AD - is an enduring tribute to one man's vision.

was driving his car. The haunting beauty of the young woman sta
them for the sweet scent and haunting beauty of their flowers. To a
Days. Her voice retained its haunting edge, and when she reached fo
cold in his body. There was a haunting feeling of familiarity in the
e fought, in Matthew Arnold's haunting image, on a darkling plain sw
Aztecs. Everything else - the haunting keyboard and nagging soprano
ed in black lace, and wails a haunting lament similar to Ofro Haza,
useums. We'll see the craggy, haunting land that the Berbers, an
es are part of an ancient and haunting landscape, and it is the livi
etry of his music has its own haunting lilt, vocabulary and rhythm.
ches, and listen to fado, the haunting music so expressive of the
d have-not society. This is a haunting novel that should give John M
; 14.99) quickly turns into a haunting parable of our times. There i
d it contains a sensitive and haunting performance from Rade Serbedz
all restrictions. Wistful and haunting piano music by Erik Satie;
and Demi Moore danced to the haunting record in the film Ghost - th
t imperious, with a dazzling, haunting smile; but the performance is
ntingly sung her own, quietly haunting song. Ex-S A Far Cry from
t surely have appreciated the haunting sound of the pipes after 280
Prevert, Francois Dupeyron's haunting tale of a husband, his wife a
ter still, in Luke's fragile, haunting voice, his effortless melodic

theft, damage to machinery or persistent absenteeism, and the employ
for just 27 runs. Apart from persistent abuse directed at home capt
from any body opening, any persistent change in a wart or mole -
of Iraqi government. Iraq's persistent claim is that the allies' a
diness when confronted with a persistent condition such as traumatic
n Wilson of our Science Unit. Persistent fatigue is the fourth most
e distressed by her husband's persistent friendship with Diana, whic
elay of at least five days. A persistent front of high pressure over
ll; If you have suffered from persistent indigestion or chest pains,

130 by the unpopular poll tax and persistent inflation. At the Rome summ
131 freelance scholars. Yet the persistent popularity of the subject i
132 However, if memory loss is a persistent problem, there are exercise
133 mic sound of the train sets a persistent pulse that throws the
134 good, but no more. Under more persistent questioning he admitted tha
135 eam against the ebb tide. The persistent rain had made the river ang
136 had his prayers answered with persistent rain over the last 48 hours
137 economic reinvestment and the persistent recession, while Perot can
138 courts", and about its persistent rejection of international
139 with relish. Yet there is a persistent risk in using these snails.
140 [caption] Slow growth and persistent unemployment are global pro