Putting frequencies in the dictionary

Adam Kilgarriff*
Information Technology Research Institute
University of Brighton
Brighton BN2 3FR
email: Adam.Kilgarriff@itri.bton.ac.uk

May 1996

Abstract

A central fact about a word is how common it is. The information is particularly valuable for language learners, as it immediately indicates how important it is to learn a word. With the advent of large computerised language corpora, it is for the first time possible to meet the demand. Both Longman Dictionaries and Collins COBUILD decided to present frequency information explicitly in new editions of their learners' dictionaries. The paper describes how this was done at Longman, and the various issues encountered along the way. It also compares the Longman and Collins COBUILD lists.

1 Introduction

A central fact about a word is how common it is. The more common it is, the more important it is to know it. All else being equal, more common words should be taught to foreign learners first, both so that they understand them and so that they know how to, and are inclined to, use them. Also, the more common a word is, the more likely it is to exhibit polysemy, irregular morphology and other idiosyncratic behaviour. Most dictionaries, and learners' dictionaries in particular, tend to say more about commoner words, so column inches provide a roundabout indicator of word frequency (Kilgarriff, 1994). But with the advent of large computerised language corpora, we can do far better. We can explicitly state which words are the common ones. Two recently-published dictionaries, Longman Dictionary of Contemporary English (3rd edition; hereafter LDOCE3) and the Collins COBUILD English Dictionary (New edition) do just this. This paper reports on how it was done at Longman, and compares the frequency information in the two dictionaries.

Della Summers, Director of the LDOCE3 project, decided to mark, in the dictionary, all words that were amongst the top 3,000 items in either spoken or written English.

*The work was undertaken at Longman Dictionaries and at Brighton University where the author was supported by EPSRC grant GR/K/18931

For each variety, we would mark the first, second and third thousands separately, using symbols S1, S2 or S3 for spoken and W1, W2 and W3 for written (a set I call 'SW-symbols' below). So if *ability* is marked $\boxed{\textbf{W1 S2}}$ this means it is in the top thousand items in written English and the second thousand items in spoken English, and if *bungalow* is marked $\boxed{\textbf{S3}}$, it is in the third thousand in spoken English (and not in the top 3,000 at all in written English).[1]

The paper will be of interest both to others considering a similar task (and for them we provide details of the problems we encountered in the sections marked *) and to the more general reader, as frequency considerations cast new light on old problems, and raise new questions which are likely to become recurring themes in the dawning age of corpus lexicography.

1.1 Structure of the paper

The task is essentially very simple:

- take a corpus
- extract a frequency list
- compare it with the dictionary to identify and rectify mismatches
- identify the one-, two-, and three- thousand cut-off points
- mark the corresponding dictionary entries accordingly.

The paper follows a similar route. Following some preliminary remarks, we address the central question of corpus composition and briefly describe our corpora (section 2) and how frequency lists are generated from them (section 3). We then describe the dictionary (section 4) and various kinds of special case which came up when comparing it with the frequency lists (sections 5 and 6). In section 7 we consider issues relating to multi-word items, and make some observations about their frequency distributions. In section 8 we describe some improvements we were able to make to the LDOCE3 text as a result of the exercise. Finally, we assess the validity of the lists we produced, including a comparison of Longman and COBUILD lists, and the results of an experiment comparing two halves of a corpus with each other.

1.2 Preliminaries 1: Zipf's Law

In a Zipfian distribution, the most common item has twice as many occurrences as the second most common, three times as many as the third, a hundred times as many as the hundredth, a thousand times as many as the thousandth, and a million times as many as the millionth. Zipf observed as long as 60 years ago that word frequencies follow this distribution (Kipf, 1935).[2] We mention it here because it has implications for almost all aspects of a word-counting exercise. It means that there are great disparities between frequencies of more and less common words, and the most common words of all –*the, be* (in all its forms) *of, and, a* – are several orders of magnitude more frequent than most other words. *The* is the most common word in the language, accounting for one word in fifteen, with a count of 6.2 million in the British National Corpus (BNC). Thus *the* is a thousand times more common than a common word like *district*, which is itself a hundred times more common than a word which is by no means obscure, such as *sunburn*.

The Zipfian theme echoes through the paper (see particularly sections 2.4, 3.2, 7 and the final discussion).

1.3 Preliminaries 2: A computer-assisted, divide-and-rule methodology

The exercise was undertaken in a 'computer-assisted' manner. It is a striking feature of the lexicon that, most of the time, most words behave regularly, but there is a small class of special cases (Jackendoff, 1975). There were many exception-classes relevant to the practicalities of our task, such as irregular verb forms occurring in the dictionary headword list (e.g. *flown, flew*), or words with different British and American spellings.

To apply some operation to 3,000 items by hand always takes a long time. If the 95% of regular cases can be differentiated by computer from the 5% irregulars, then the 95% can be dealt with automatically and the 5% looked at more closely. The 5% can often, itself, be subdivided into regulars and irregulars, with another program being written to deal with the "regular irregulars", and so on iteratively. The list of 3,000 is rapidly decomposed into a small number of types, each of which can be identified and dealt with by program, and a small residue of anomalies, which can quickly be dealt with by hand. We used this approach many times over in the course of compiling the lists. Without it (and a set of programming tools which allowed quick searching, sorting, and editing operations) the exercise would have been far slower and more tedious.

2 Corpus composition

2.1 The representativeness problem

There is one major theoretical problem with the task. For the frequencies to be true of general English, they would have to be taken from a representative corpus of general English. But there are many obstacles to making this a coherent concept[3] (Biber, 1993). What should be in such a corpus? The unanswerable questions include: what should the ratio of written to spoken material be? Is it the amount that is spoken or written which is important, or the amount that is read or heard (so does a copy of The Sun count for more than an obscure treatise on dog whelks)? What, in terms of dialects, date (of speaking or writing or hearing or reading), and for non-native speakers, competence-levels, demarcates the 'general English' we wish to represent.

These theoretical questions resolve into the following more manageable ones: what different text-types do we want in the corpus, and in what ratios? How many samples of each, and in what sample sizes? And the entirely practical question that this then collides with is: how many samples of each text type can we get hold of, with what copyright restrictions, and at what cost. The outcome is a corpus which will never be beyond challenge at a theoretical level, but which does nevertheless allow us to address with a degree of objectivity some central questions about the language, where before we could only speculate.

We mention these issues here although they are not the subject of this paper because they underlie the entire exercise. The quality of the frequency list, and thus the validity of all that follows, is premised on the composition and representativeness of the corpora.

2.2 Corpus resources

The main corpus we used was the British National Corpus (BNC), a resource developed by a consortium whose members were Oxford University Press, Longman Group Ltd, Chambers Harrap, Unit for Computer Research on the English Language (Lancaster University), Oxford University Computing Service and the British Library Research and Development Department.[4] It comprises 90M words written English and 10M words transcribed spoken English. In a secondary process we used the American English component of the Longman Lancaster Corpus[5] (Summers, 1992) and the Longman Corpus of Spoken American English, to run a series of checks to ensure that words common in American English were not overlooked.

2.3 The two parts of the spoken BNC

The spoken BNC caused us a particular problem. It has two halves: the 'demographic' half is everyday, natural conversations between families, friends and so forth, and the 'context-governed' half consists of slightly more structured spoken language in meetings, lectures and broadcasts (Crowdy, 1993). To begin with, we worked from a list which had been generated from the whole corpus. But as we proceeded, we became concerned about the high prevalence of words like *audit, auditor, committee, motion.* The lexis of meetings was playing a dominant role. So we subjected the list to a further edit, on the basis of frequencies in the demographic half alone.

2.4 The whelks problem

We also checked for words where all the occurrences were from one or two corpus samples. If a book is about whelks, it might well use the word *whelk* several hundred times, so a sample taken from that book might contain over a hundred occurrences for a word that, outside that context, is rare. The problem can be countered by taking very many sources, and keeping the sample size low, and this was the strategy adopted in principle in the BNC.

However, in practice, collecting more, smaller samples was far more costly than collecting fewer, larger samples and the eventual outcome was a compromise. Where the source was a book, the standard sample size was 40,000 words (so a single source would account for 0.04% of the corpus, as against 0.2% in the LOB and Brown corpora). For other types of source, the issue was more complex since a single source could include a variety of text types. For example, a newspaper contains leaders, sports journalism, and personal advertisements; also, copies of the same newspaper of different dates are, at a first pass, different documents even though they share the same source. Six sources accounted for more than one million words (1% of the corpus) each, but four of these were newspapers, one was a news service, and one was Hansard, the official record of the UK Parliament. None of these consistently used specialised vocabulary, so they did not give rise to the whelks problem. The one source that did to a substantial extent was *GUT: Journal of Gastroenterology and Hepatology*, which accounted for 713,000 words (0.7%) of the corpus. Words such as *peptide* and *endoscopy* needed weeding out of the frequency list.

There was a further complication in that a single source did not correspond to a single BNC file (which currently corresponds to a BNC 'document'). How text of a single source might be distributed across files, where it became too large for a single file, or changed in subject matter, text type, authorship, edition (of a journal or newspaper), or in some other way, has been an ongoing problem in the development of the BNC.

The written BNC had enough sources so that the whelks problem arose rarely, but this was not always true for the spoken BNC or the American corpora. Additional checking was required to ensure that words were not included in the top 3,000 because of their frequency in just one or two sources.

3 Generating a lemmatised frequency list

The BNC was loaded into CorpusBench[6], a corpus database system. CorpusBench, when armed with an inflections dictionary and looking at a part-of-speech(POS)-tagged corpus, is able to provide a lemmatised frequency list. The entire BNC has been POS-tagged by the CLAWS system (Garside, 1987; Leech, Garside, & Bryant, 1994), and we were able to generate an inflection dictionary from existing Longman resources. By 'lemmatised', we mean two things. First, for verbal *aim*, the count will consider all instances of *aim, aims, aiming, aimed*; and second, it will exclude all non-verbal instances, so nominal *aim* and *aims* will not be counted. The count will be of verbal instances only of any of the four forms.

3.1 Lemmatisation details *

For CorpusBench to make use of this facility, a substantial amount of machinery is required. The POS-tags in the corpus must be in two parts, the first part being the major word 'class', (eg. *noun*), the second being the 'form' (eg. *singular* or *plural*). After POS-tagging, the corpus input

```
fat porcupines
```

becomes, in the dialect of SGML required by CorpusBench,

```
<W c=adj f=base>fat</W> <W c=noun f=plural>porcupines</W>
```

where `W` is the symbol for the 'word' element, `c` for the 'class' attribute and `f` for the 'form' attribute, and `base` shows that we have, e.g. the base form of the adjective.[7]

The inflection dictionary states, in its preamble, which forms are possible for each class, and then, for each word, what its class and inflected forms are. So the preamble states that nouns have singular and plural forms, and the entry for porcupine states that it is a noun, with the singular inflected-form porcupine and plural inflected-form porcupines. There may also be zero, or several, inflected-forms; information has no inflected-form for f=plural and corpus has two, corpuses and corpora. Where a <word, class, form> triple in the corpus does not tally with the inflections dictionary, it is counted as a separate object. Where a triple occurs twice in the dictionary, an occurrence of the triple in the corpus will be double-counted, as it will contribute to the scores of all the lemmas where it belongs; hence any corpus occurrences of *ring* as the base form of a verb will add to the scores of both *ring-rang-rung* and *ring-ringed-ringed*.

3.2 Frequency list: first-pass edit *

The resulting frequency list, when run on the written part of the BNC, contains over a million items.[8] There are two kinds of items; those that correspond to words in the inflections dictionary, and those that do not. The former are <word, class> pairs, the latter <word, class, form> triples. Amongst the latter were many proper names and all closed-class items. (Our inflections dictionary contained only nouns, verbs, adjectives and adverbs. As these are the only word classes taking inflections in English, there was nothing to be gained by including the others.)

The first thing to do was to reduce the list to a manageable length, which we did by throwing out all items which were well outside the top 3,000. We set a threshold of 2,000 occurrences. This gave us 3,679 items. At a stroke, this eliminated many of the problems relating to anomalies such as mis-spellings which can be such a bugbear for people working with wordlists: anomalies were generally not repeated 2,000 times so were deleted at a stroke.[9]

Letters of the alphabet were deleted from the list on the basis of their POS-tag. So were proper names. This was the one move which threw out a substantial number of items: 240, out of the list of 3,679. *Mark* may be a verb, a common noun or a proper noun. CLAWS made a choice for each occurrence, and all three ended up in the top 3,000. We deleted the proper noun leaving the verb and common noun. As well as straightforward names, placenames (*road, street, north, college*) and titles (*general, lord, lady*) were deleted in their proper-name, though not their common-noun, readings. (All these items occurred in the top 3,000 list in both readings.) We did not use capitalisation in any of these decisions because CLAWS had already made use of the information implicit in capitalisation in its choice of tag, so it would be giving it undue weight to consider it again.

4 The dictionary

The Longman Dictionary of Contemporary English, 3rd edition (LDOCE, 1995) is a new corpus-based dictionary of English for foreign learners. It has quite a number of new features, including graphs showing relative frequencies of words, collocations and grammatical patterns across spoken and written corpora: here I shall only consider those that relate to the labelling with S-W symbols.

The editorial decision was that SW-symbols would be placed in the margin alongside entries. Thus the question, "what is an LDOCE3 entry?" was crucial. In all cases, that was what an item in our frequency list had to match up with. LDOCE3 policy was that an entry corresponded to a <word, class> pair; a word-class ambiguity always gave two homographs, but a semantic ambiguity alone never did; *bank* has two homographs, one for the noun, the other for the verb. The 'money bank' and 'river bank' meanings are different senses at the noun homograph.

This was very fortunate, since, for the great majority of cases, it meant LDOCE3 headwords corresponded directly to items on the lemmatised frequency list.

One other salient LDOCE3 policy related to ordering of homographs. The policy was that they were ordered by frequency. The frequency-list exercise meant that the lexicographers' decisions could be re-checked in a systematic way.

4.1 Two lists and a dictionary: running a comparison

The two lists (one for written English, one for spoken) needed amending so that they only contained <word, class> pairs corresponding exactly to LDOCE3 entries. This was done with LDOCE3 on one computer window, the BNC in CorpusBench in another, alphabetically-ordered versions of the two lists in two more, and the output of a program which took the lists and LDOCE3 as inputs, and gave any mismatches or mis-orderings as output, in another. This rapidly threw up the various questions that occupy the following three sections.

5 Part-of-speech discrepancies

No two linguists will come up with exactly the same list of 'the word classes of English', and so it is scarcely surprising that some CLAWS categories[10] have no exact match amongst the LDOCE3 parts-of-speech. It is fortunate that both LDOCE3 and CLAWS are descendants of the analysis in (Quirk, Greenbaum, Leech, & Svartvik, 1972); this certainly reduced the degree of discrepancy. The problems arose with the minor parts of speech: predeterminer, quantifier etc. For words with one or more of these part-of-speech labels, the LDOCE3 entry or entries needed checking against the corpus data to determine the appropriate match.

5.1 Many-to-one and one-to-many POS mappings *

There were, inevitably, various exceptions to the 'one word class per entry' rule. *Determiner, pronoun* is an acceptable LDOCE3 part-of-speech for those words (e.g. *any, both, each*) where the distinction would not be helpful to a language learner. In such cases, if only the determiner, or only the pronoun, was in the list, all that needed doing was amending the item in the frequency list so that the part of speech read *determiner, pronoun.* Where both were present in the list, as in

45180	each	determiner
9453	each	pronoun

this needed amending to

54633	each	determiner, pronoun

We amend the number, in this case by adding the scores for the two items together, as the numbers would determine whether the LDOCE3 entry was classified as a W1, W2 or W3 item.[11]

The many-to-many and one-to-many mappings more problematic in principle, but less so in practice. An example: the verb *do* only had one entry in the frequency list but two, one for the auxiliary and one for the lexical verb, in the dictionary. In practice, for all the items where this occurred, it was evident that both forms were extremely common, so

356792	do	v

became

178396	do	v
178396	do	auxiliary verb

There may be no justification for assuming the two forms are equally common, but since they were both going to be 'W1 S1' items in any case, the imperfection would make no difference. As it turned out, this applied wherever there was a many-to-many or one-to-many mapping. A potential problem never came to pass, owing to the strong correlation between frequency and idiosyncratic behaviour.

6 The wordishness of words

There were various 'words' in the dictionary which matched 'words' in the frequency list, but which, nevertheless, we decided not to include amongst the frequent words, because they were insufficiently word-like. These included numbers, closed sets such as nationalities and currencies, non-standard forms, and variants.

6.1 Numbers *

Quite a few numbers were in the list. Numerals do not have entries in LDOCE3 and, in written English, are clearly not words so it is neither appropriate not possible to include them. Numbers when spelt out (e.g. *three* rather than *3*) were a less straightforward case, but again, we decided to exclude them. The primary reasons were pedagogical. The information was very unlikely to be of any interest to the dictionary user, and the differences between different numbers would look peculiar. There was also a technical one. While the distinction between numeral and spelt form is straightforward for the written corpus, it is non-existent for the spoken corpus. The instructions for transcription of the spoken material stated that all numbers should be spelt out (Crowdy, 1994). Hence, as an artifact of the transcription scheme, spelt forms of numbers were generally more common in the spoken than the written corpus. The only way to treat the two in a consistent manner was by excluding all numbers.

6.2 Closed sets *

Days of the week, months, currencies, countries, nationalities, religions (the last three in nominal and adjectival forms) all occurred in the top 3,000 list. We decided that these items, though wordlike enough to be in the dictionary, were not wordlike enough to count for the purposes of the frequency list. Syntactically, they operate as proper names, and indeed, many had been deleted when proper names had been deleted from the lists; orthographically, they are often or always capitalised; pedagogically, they are not the sort of item where frequency information is of any interest.

6.3 Interjections and non-standard forms *

At one end of the spectrum was *hmm*, at the other was *gosh*. Interjections were, naturally, particularly prominent in the spoken corpus, and the particularities of the transcription scheme played a substantial role in determining what was in the top 3,000. The transcribers had been instructed to transcribe "orthographically", that is, by giving words their standard English spellings (as opposed to doing anything to mimic the sound of the word on the audio tape). To cover those common items which had no standard English spellings (and those cases where the distinction between spoken form and standard form was sufficiently marked, e.g. *gonna*), the transcribers were given a control list of slang and dialect forms from *attaboy, botty* and *brekkypoohs* to *wotcher, wibbly* and *yuk*, and another for vocalised pauses from *aargh, ach, ah* to *urgh, weeoow, whoop*.[12]

Those items such as *urgh* which had no LDOCE3 entry could not have an S1, S2 or S3 put beside them so were deleted from the list. Of those which did have LDOCE3 entries, a judgement was made of how word-like they were: *gosh, dear* and *hi* were left in but *ah, er* and *um* were taken out.

6.4 Cross-reference entries and spelling variants *

There were a couple of minor questions here. Both *colour* and *color* are common items, but they both relate to a single word. LDOCE3 makes headwords of both of them, but, at *color* the user is redirected to *colour* where there is a full entry. We chose to treat the variable-spelling word as one item, adding together the counts for both forms, and counting it as one item in the top 3,000 frequency list.

Irregular verb forms and plurals (*is, done, broken, children*) generally did not feature in the frequency list, as their frequencies were counted under the base form of the word (*be, do, break, child*). We did not put SW-symbols at the cross-reference entries for irregular forms.

7 Multiwords

In the (written) language there are various kinds of word-like items which are spelt with spaces between the constituents. Different theorists have different lists. We shall call them multiwords. The overlap between what LDOCE3 treated as a multiword and what CLAWS did was virtually nil.

A computer wordcount program defines a word as any string of characters separated by blanks or punctuation. It has no truck with multiwords. Barring minor quibbles about hyphens and apostrophes, that makes counting easy and has the advantage that everyone knows where they stand and will arrive at the same numbers. The disadvantage, of course, is that it doesn't tell the truth. Sometimes sequences of characters which include spaces should, for all sorts of good linguistic reasons, be treated as a single word.[13]

Any step towards the truth (as linguists strive to define it) tends to be a step away from anything that is computationally straightforward. First CLAWS multiwords, and then LDOCE3 multiwords, had to be reconciled with the CorpusBench-generated frequency list, as follows.

7.1 CLAWS multiwords *

On examining the output of earlier versions of CLAWS, the Lancaster team had noted that many of the errors were caused by the difficulties the program had with a small number of multiword adverbials, determiners, prepositions and conjunctions; *as well as, at least, no doubt, because of* etc. When trying to tag the three words in *as well as*, CLAWS would often make mistakes, in part because each word can take several parts of speech, in part because the whole expression operates idiosyncratically. These mistakes would also cause further mistakes in tagging the surrounding words. So they modified CLAWS so that it first looked for these items, and, when it found one, it tagged the whole unit as, possibly, a single multiword. The probabilistic part of CLAWS then considered the single-item interpretation(s) of *as well as* alongside the three-item interpretations, and chose the most probable. Its performance then improved (Leech et al., 1994).

There are around 600 of these multiword items. In the published version of the BNC, the multiwords are tagged as 'words'. Most have fairly high frequencies, so they have a substantial effect on wordcount, collocational, and other statistics.

When we came to build the BNC into CorpusBench, we had two options: keep to the CLAWS treatment, or revert to a spaces-always-separate-words philosophy. The decision would dictate the shape of our corpus resources for some time to come, so we considered it carefully. In favour of the CLAWS strategy were the arguments that it presented an accurate picture of the multiwords, and would facilitate investigations into their behaviour.

On the other side were various points relating to the transparency of the system. Since *no doubt* was a multiword, a lexicographer examining the behaviour of *doubt* and so calling up a concordance for *doubt* would not find any occurrences of *no doubt* in it. If she asked CorpusBench to list collocates of *doubt*, *no* would not be amongst them, nor would items that frequently preceded or followed *no doubt* (but did not often precede or follow 'non-no' *doubt*). Where words participate in more than one multiword, and where frequencies and mutual information scores are concerned, the results are misleading in complex and confusing ways.

Lexicographers would not take kindly to being expected to hold in their minds at all times an extra six hundred difficult items, and would rightly say that the corpus system should be helping them, not hindering them. Our primary criterion was making the corpus lexicographer-friendly, so the decision was to break all CLAWS multiwords into their constituent parts.

This had the complication that we had no viable POS information about the constituent words: if the input tells us, simply, that *as well as* is a conjunction, but we want to treat it as three words, we will not know what the POS for *as, well* or *as* is. We decided to assign the POS of the whole multiword to each constituent word, as a least bad option. So the subordinating conjunction

```
<W c=c f=s>as well as</W>
```

became

```
<W c=c f=s>as</W> <W c=c f=s>well</W> <W c=c f=s>as</W>
```

This meant that our top 3,000 frequency list contained items such as *well* as a subordinating conjunction. Some detective work was required to determine which multiword or multiwords such usages related to. Once the multiword had been identified, we sought out the LDOCE3 entry where it was defined. Then the frequency lists and counts were amended as described in the discussion of many-to-one POS mappings, section 5.1 above.

7.2 LDOCE3 Multiwords *

Whereas CLAWS multiwords were closed class items, LDOCE3 multiword headwords were almost all compound nominals. The issue here was, simply, which were common enough so that they should be in the top 3,000.

CorpusBench could only answer questions of the form "how common is this word-pair?" There was no convenient way of providing it with a list of multiwords and asking it to provide frequencies for them all.

Reasoning that the frequency of *word1 word2* cannot be higher than the frequency of whichever is rarer out of *word1* and *word2*, we took a list of all LDOCE3 multiwords, and threw out all those where one of the constituents was not in the top 3,000. We then scanned that list (which had about 2000 entries) to see what might conceivably be in the top 3,000. We checked several dozen items, but most were well below the thresholds. The only items added to the lists at this point were *ice cream*, *of course*, *all right* and *according to*.

One large and important category of multiwords is phrasal verbs. Pedagogically, it would have been desirable to mark these with SW-symbols, since many phrasal verbs are very common and are natural alternatives to non-multiword near-synonyms. In LDOCE3 phrasal verbs are nested within verb entries, but always start a new line, so it would have been typographically possible to mark them with SW-symbols in the margin. However, it was not possible to count them systematically. CLAWS does not distinguish "look for ten minutes" from "look for the key", and while a search for the nodeword *look* with the supplementary word *for* somewhere among the five following words would give a fair indication of the frequency of the phrasal verb, it would include some false positives ("look for ten minutes") and miss some instances ("look here, there and absolutely everywhere for the key"). Moreover, as noted above, CorpusBench offered no straightforward way of generating frequency lists for multiwords, and the question, "how do you count the instances of *look* if you are also counting all its phrasal verbs" was one we wished to avoid, since each possible answer to it had some undesirable implications.[14]

So we decided against putting SW-symbols beside phrasal verbs. Only main entries would carry SW-symbols. The theoretical, practical and pedagogical advantages of the simple principle outweighed the pedagogical benefits of marking multiwords.[15]

7.3 Multiword frequencies: general observations

It is only very, very common words which participate in very common multiwords. The most common multiword a word occurs in rarely accounts for more than a quarter of its occurrences. For a word's most common multiword to be over the 2,000 threshold, the word would generally be amongst the 500 most common words of all.

Almost none of the LDOCE3 multiword headwords were in the top 3,000. The nouns and adjectives that are their constituents are not generally very very common: the tops of frequency lists are dominated by closed class words. On the other hand, the CLAWS multiwords, whose constituents are mostly closed-class items, and which themselves operate as closed-class items, were often very common.

8 Checks on LDOCE3 text

The exercise served as a check on the LDOCE3 headword list in several ways, as described below.

8.1 Omissions and Run-On entries *

It revealed a small number of omissions and a number of items which were in the top 3,000 but had run-on entries only.[16] The latter were mostly –ly adverbs: *automatically, carefully, closely, rapidly*, sometimes adjectives with past participle form: *increased, expected*, and occasionally 'other': *implementation* at verbal *implement*. LDOCE3 policy was that a word should only be entered as a run-on if, amongst other things, it was less frequent than the headword and was not in itself a frequent or important word, so the run-on status accorded these items was clearly an error.

Full entries were promptly written for any run-on or omitted items in either of the top-3000 frequency lists.

8.2 Homograph re-ordering

It was LDOCE3 policy to order homographs by frequency. This had not been an easy policy for lexicographers to implement. In some cases the frequency order was evident: we all know verbal *see* is more common than the bishop's *see*, but in many it was not. For verbal versus nominal *aim, help, play* etc. our intuitions are of little use. A side-effect of the frequency-list exercise was that we were able to check homograph-ordering decisions automatically.[17] Spoken and written corpus were given equal weight, and where the LDOCE3 ordering did not tally with the corpus scores, LDOCE3 was re-ordered.

9 Reliability

How good is the list? Is it 'true'?

Let us first rework the question according to the scientists' favourite criterion: replicability. Would another team, working in the same framework, with the same goals, arrive at the same list? (We take 'the list' to be a list containing all the items from the two top 3,000 lists, marked with the SW-symbol(s) they get in LDOCE3.) This then breaks down into two questions: firstly, would it be the same list if they used the same corpora, and secondly, how similar would it be if they had not.

9.1 Replicability using the BNC

There is a multiplicity of reasons why the list would not be exactly the same even if the same corpora were used. Firstly, there are all the special cases. Different choices would have been taken on numbers, closed sets, cross-reference items and interjections; not all decisions (or sums) relating to part-of-speech discrepancies and CLAWS multiwords would have been the same; the threshold for the 'candidate list' which was then pruned would not have been the same, resulting in further minor arithmetic differences. But each of these sources would only result in a very small number of differences to the list, particularly as there are only six SW-symbols: small differences in the number associated with an LDOCE3 entry would not usually change the symbols it was allotted. Only if the difference between numbers meant that the item crossed a threshold, switching, say, from "S2" to "S1", would the list be different. We think it unlikely that these sources of anomaly would affect more than 2% of the items on the list.

Secondly, our main corpus was the BNC – but we did not want our list to directly reflect the BNC because, firstly, in its spoken component, meetings were over-represented (see section 2.3), and secondly, it covered only British English. Our list was also to cover the American variety. Our American corpora were smaller, so intrinsically a less dependable source of counts, and there was no objective way to "add" the American and British lists together. The process of compensating for features of the BNC was, inevitably, subjective; deleting meeting-ese, and adding in high-frequency words from the American corpora where an American lexicographer with a good knowledge of British English assessed that the reason they were not already in the BNC list was that they were more common in American. Around 200 items of meeting-ese were deleted and 100 American words added, mostly to the S3 and W3 categories.

So while there were many reasons why another team working on the BNC would not have had exactly the same list, their list would not be very different to ours. As noted in section 1.3, most words are regular, and most of the irregularities fall into classes. It was generally only the irregular irregulars which prevented the pure-BNC list from being entirely replicable, and these were few in number. At the point where we started looking outside the BNC, this ceased to hold.

		Rank order in half-1						
		1-500	501-1000	1001-1500	1501-2000	2001-2500	2501-3000	over 3000
Rank	1-500	**482**	17	1	0	0	0	0
order	501-1000	18	**440**	39	2	0	1	0
in	1001-1500	0	43	**400**	51	3	0	3
half-2	1501-2000	0	0	56	**356**	79	5	4
	2001-2500	0	0	4	79	**314**	87	16
	2501-3000	0	0	0	9	81	**303**	107
	over 3000	0	0	0	3	23	104	

Table 1: Similarity between rank order of most common words in random-half subcorpora of the BNC. If the rank order of words had been identical for the two half-corpora, then all the numbers on the leading diagonal (from top left to bottom right, in bold) would have been 500, with all other cells being zero. The first row states that, of the 500 most common words in half-2, 482 were amongst the top 500 in half-1, 17 were amongst the second 500 (e.g., having ranks between 501 and 1000), and one was in the third (ranks 1001-1500).

9.2 Replicability using a different corpus

Had the other research team used a different corpus – of comparable size and similar design, but containing different texts – their list would certainly have been different. We performed an experiment with the BNC to determine how different it might be. We divided the BNC into two halves by randomly assigning each of the 4124 documents in it to one of two subcorpora, "half-1" and "half-2". We then compared the frequency lists for the two halves - that is, for two 50M word corpora, gathered using identical criteria and policies and tagged using the same tagger. The results are presented in Table 1.

For these two corpora, only 130 words in the top-3000 lists were different. No words were in the top 1,000 in one half, but outside the top 2,000 in the other, and only 10 words were in the top 2,000 in one half but outside the top 3,000 in the other.

The table shows that two lists for different corpora will show a higher proportion of top-500 items in common, than for second-500, and so on down. This can be seen from the leading diagonal of the table: as we progress down the table, the number of matches deviates further and further from 500. This is a consequence of at least three factors; first, changes to the W1 items have knock-on effects on what is a W2 or W3 item, but not *vice versa*. Second, it is a general statistical truth that claims based on bigger counts are more reliable. The count for *the* is five orders of magnitude bigger than that for *sunburn*, so we can be far more confident that *the* occurs approximately once in every 15 words in general English, than that *sunburn* occurs once in every 1.5 million. It would be very surprising to find a corpus where the frequency of *the* halved, down to once in every 30 words, but not to find one where the proportion of *sunburn*'s halved, falling to one in 3 million.[18] Thirdly, the Zipfian distribution means that most W1 words are far above the W1 threshold, but most W3 words are quite near the W3 threshold. The word halfway

No. filled diamonds	Frequency rank	Number in band
5	1–700	700
4	700–1,900	1,200
3	1,900–3,400	1,500
2	3,400–6,600	3,200
1	6,600–14,700	8,100

Table 2: COBUILD2 frequency bands

through the W1 class – the 500th most common word – is nominal *trade*, with a count of 17,942. This is exactly double the W1 threshold of 8,935. But the word halfway through the W3 class, the 2,500th most common word, is *cake*, with 2,944, just 32% above the W3 threshold of 2229. This compounds the second reason; lower counts are intrinsically less reliable, so more likely to vary between corpora, and then, if the counts for a word are different between two corpora, the lower the counts, the greater the effect will be on the word's position in the frequency lists for the two corpora. Taken together, this means that it would be surprising to find a comparable corpus where *trade* was not W1, but not to find one where *cake* was not W3.

The four-way categorisation (W1, W2, W3, no-W-symbol) acts to our benefit. It makes clear that we are not pretending to make very fine-grained frequency claims. Also, the W1 symbol, which is in any case the W-symbol that the dictionary-user will take most note of simply because it relates to the most common words, is the symbol built on firmest ground.

9.3 Comparison with COBUILD

The 1995 edition of the Collins COBUILD English Language Dictionary presents frequency information in a manner similar to LDOCE3, so provides an opportunity for direct comparison. (The published dictionary was our only source of information on the COBUILD approach, so a full comparison between the methods, as opposed to results, was not possible.) The comparison revealed differences on various fronts, as detailed below. Yet despite that, when the markings for the top three thousand words in the two dictionaries were compared, a high degree of overlap was found.

In contrast to the Longman S-W symbols, COBUILD uses a system of filled and unfilled 'diamonds', with all five diamonds being filled if the word is among the 700 most frequent words in the language, down to just one being filled if it is between 6,600 and 14,700 in the frequency list. COBUILD does not distinguish written and spoken frequencies. The relation between diamonds and frequency-bands is given in the front matter, and is as shown in Table 2.

As in LDOCE, the information is given in a column alongside the text for the relevant entry, so COBUILD diamonds are constrained to correspond to COBUILD entries as Longman's SW-symbols are to LDOCE entries. This makes for another marked difference between the lists. COBUILD entries are for words irrespective of part of speech, so, e.g., nominal and verbal *pass* are addressed within the same entry, whereas in LDOCE they form two different entries.[19] The COBUILD approach still requires its text to be part-of-speech tagged if it is to be accurate, since it does need to distinguish between, e.g., *frank* as a name or as an adjective, or *acting* as present participle of the verb (in which case it contributes to the count for the *act* headword) and as the noun. Whether to distinguish headwords on the basis of part of speech is a general lexicographic question, rather than one specific to frequencies, so we do not discuss it further here.

COBUILD's claims cover far more of the vocabulary than Longman's, and for that reason, are less reliable, as discussed above. Taking a page of COBUILD, words with one diamond are *abandonment, abate, abdomen, abdominal, abduct, aberration, abide* and *abiding* whereas those with none include *abashed, abatement, abattoir, abbot, abbreviate, abbreviation, abdicate* and *abhorrent.* These two lists could be swapped over without the results being in the least at odds with our intuitions about the language. The evidence supports the hypothesis of the previous section: even on the basis of very large corpora, word counts for all but the few thousand most frequent words are highly unstable.

As discussed above, LDOCE set aside closed sets, numbers, and all items which, as headwords, were capitalised. Closed sets like the signs of the zodiac, which were well outside the top 3,000 so did not require consideration at all at Longman, were all given two diamonds in COBUILD. In relation to semi-closed sets such as nationalities, where a same-frequency assumption is clearly invalid, this approach is awkward. COBUILD gives each set member, from *British* and *American* to *Yemeni* and *Bolivian* four diamonds. Numbers are marked and are not treated as a set, so, for example, *thirteen, fourteen, seventeen* and *eighteen* have five diamonds, *fifteen* and *nineteen*, four and *sixteen* two.

COBUILD have always insisted that it is impossible to create a corpus that is truly representative of the language, and have focussed on size of corpus rather than balance. Their frequency lists are based on the 200 million word "Bank of English". It includes a higher proportion of newspaper and newswire material than the BNC, as is evident from the frequency ratings for *newscaster* (five diamonds), *accord* (five), *correspondent* (four) and *sanction* (four). These words were all well below the threshold for any SW-symbol in the BNC.

The Longman list is able to present information that COBUILD misses because it addresses written and spoken language separately. *Extensive*, for example, is a three-diamond word in COBUILD, but is in fact only common in written English, so the implication for the learner wanting to speak English is misleading. In LDOCE it is marked W3, with no S-symbol.

But, these differences notwithstanding, how great a correspondence is there between COBUILD and LDOCE lists? One might expect an approximate match between 5-diamond and S1 and/or W1 words, 4-diamond and S/W2 words, and 3-diamond and S/W3 words. (Many COBUILD headwords have multiple POS's so map to multiple LDOCE headwords, so the COBUILD top seven hundred will map to a larger number of LDOCE headwords.) For a sample of 100 words marked with 3, 4 or 5 diamonds in COBUILD, we investigated whether this did in fact hold. The results are shown in Table 3.[20]

As in Table 1, a high degree of correspondence is associated with high scores on the leading diagonal (in bold) and low elsewhere. The first row of the table states that, of the sample of 20 words marked with five diamonds in COBUILD, 18 were S/W1 words

		LDOCE score			
		1 top 999	2 1000-1999	3 2000-2999	Unmarked
COBUILD	5 (top 699)	**18**	1	0	1
score, in	4 (700-1899)	8	**19**	1	2
'diamonds'	3 (1800-3400)	1	17	**25**	7

Table 3: Match between words in LDOCE and COBUILD frequency bands.

in LDOCE, one was an S/W2 word, and one received no SW-symbol.

The table shows a close match, despite the differences between LDOCE and COBUILD corpora and policies. This strongly implies that the frequency lists as used by both publishers do have a high degree of validity for the top three thousand words.

9.4 Conclusion

A high-quality frequency list for English was produced and incorporated into a learner's dictionary. We are confident that the list contains much of pedagogical and general linguistic value. If the Longman list was compared with another, based on another large corpus built with representativeness in mind but containing different documents, there would be some change to the W3 and S3 items but very little change to the W1 and S1 items.

As well as producing the list, the exercise has started charting the territory of the frequency distributions of multi-word items – a territory in need of mapping if lexicography is to fully embrace the opportunities offered by very large corpora.

Acknowledgements

I would like to thank Steve Crowdy, Adam Gadsby, Mark Lauer, Prof. Geoffrey Leech, Della Summers, and the two IJL reviewers, Sue Atkins and Gerald Nelson, for their comments on earlier drafts of this paper.

Notes

[1]The questions "how should frequency information be displayed in the dictionary?" and "how can we teach the user to interpret and exploit it?" are both problematic and important, but lie outside the scope of this paper.

[2]Various mathematicians and linguists, from Zipf onward, have tried to explain why this might be, with inconclusive results (Baayen, 1993).

[3]As aired at, for example, a debate on the theme at the Fifth Conference of the UW Centre for the New OED and Text Research, Oxford 1989 between Prof. John Sinclair and Prof. Sir Randolph Quirk; see also (Church & Mercer, 1993).

[4]Documentation for the corpus is availability at WorldWideWeb site `http://info.ox.ac.uk/bnc`.

[5]Also available for research from Longman Dictionaries.

[6]A product of Textware A/S, Copenhagen.

[7]The conversion of the data into the form required by CorpusBench was undertaken by George Demetriou.

[8]Where we do not explicitly state a corpus or a list, we shall be referring to the written BNC; the list was generated at a point when there were 84 million words in the database so all numbers will be based on this population.

[9]The exceptions are indicative of the Zipfian theme. An error in the processing meant that, occasionally, the closing-word tag, `</W>` was placed after the inter-word space rather than before it. The error occurred roughly once in every thousand words. Where a word occurred more than 2 million times in the corpus, an error applying to one thousandth of its instances would result in a triple which occurred more than two thousand times. Correspondingly, amongst the items in the unedited top 3,000 list are five of the six most common words (eg excluding *be*, which has the morphological complication) in a variant where the 'word' includes the following space.

[10]Or, to be precise, CLAWS-derived categories; in the conversion of the BNC data into a format suitable for CorpusBench, the CLAWS tags had already been renamed and restructured into 'class' and 'form' as described above.

[11]Had the pronoun fallen under the 2,000 threshold, it would not have been in the list, so we would not have added anything to the figure for the determiner in turning it into a *determiner, pronoun* item, even though there might have been 1,999 further occurrences. We defend this as a minor sin, which will not radically effect our four-way classification of frequency (see also section 9).

[12]The control list in general increased the likelihood that interjections would count amongst the top 3,000, since, rather than *uurgh* scoring 50 and *urgh* scoring 100, these would both be spelt *urgh* which would then score 150.

[13]One particularly acute reason is that there are various 'words' which might be spelt with, or without, a space in the middle, e.g., *at least/atleast, green belt/greenbelt.*

[14]The options would be

- Leave the count for *look* as it was, that is, inclusive of all items also counted amongst phrasal verbs. This would show some correspondence to the nesting of the LDOCE3 entry, but would be counter-intuitive in its double counting. Verbs which were only used as part of a phrasal verb would get SW-symbols both at their single-word entry and at their phrasal subentry.
- The count for *look* could have all counts for phrasal verbs subtracted from it. This would be counterintuitive where a verb which scored well over (say) the S3 threshold was nonetheless not marked with S3, and nor were any of its phrasal verbs, since the hits for the phrasal verb were distributed between its simple form and its phrasal forms, with none, alone, passing the threshold.
- The count for a phrasal verb could have subtracted from it the counts for all of its above-threshold phrasal verbs. While escaping the more acute cases of problems caused by the options above, this course was awkward to implement, not only due to the built-in difficulties of counting phrasal verbs, but also because the threshold will keep changing: the threshold is not a set number of, say, 2,300 hits, but is defined as the frequency of the three-thousandth most frequent word. So each time a new phrasal verb was spotted which fell inside the threshold so its corresponding simple-verb count was reduced, possibly thereby falling below a threshold, the 3,000th-item threshold would change and the whole process would need repeating.

[15]The argument applied to other classes of multiwords, such as those that CLAWS identified, as well as phrasal verbs.

[16]A run-on entry is lexicographer's jargon for a word morphologically derived from a dictionary headword which is not itself defined, but is printed, along with its word class and possibly an example, usually in a bold typeface, at the end of the entry for the word it is derived from.

[17]It was also necessary if the dictionary was not to look foolish. Without the SW-symbols, mis-ordering of homographs might be an error, but it was a near invisible error. But with the SW-symbols, if a first homograph did not have a symbol where a second did, the error was evident for all to see.

In fact some adjustments were needed to make sure this situation did not arise, since a homograph that was just below both S3 and W3 thresholds could have a higher joint score than its co-homograph which was slightly more common in the one text type but substantially rarer in the other.

There were still some cases where the first homograph did not receive an SW-symbol but another did. For example the past participle of *find*, is, appropriately, the first *found* homograph as it is the commonest, but since it is a cross-reference entry for an irregular form, it did not get an SW-symbol, whereas the second homograph (as in "founded in 1844") did. These cases we lived with.

[18]If each word in the corpus was randomly drawn from a wider population and each word-selection was independent of all others, on the evidence from the corpus we could be 95% confident that the proportion of a word in the population at large was between

$$p \pm 1.96\frac{p(1-p)}{\sqrt{n}}$$

where n is the number of hits in the sample, p is the proportion of hits in the sample (e.g. n divided by corpus-size) and the number of hits in the corpus ($n \times p$) is greater than 5. The independence assumption is untrue: words are not randomly selected (Church & Gale, 1995; Kilgarriff & Salkie, 1996) but it remains a fair approximation that the confidence interval increases with the inverse of the square root of the count, which we adopt for illustrative purposes.

We has 233,921 hits in 84 million, so *we* occurs 2,790 times per million, whereas *privilege* is one hundredth as common with 2,338 hits, or 27.9 per million. The square root of 100 is 10, so we can then be as confident that the actual proportion of *we* in general English (or within another comparable corpus) is within 200 of 2,790 for *we* as we can that, for *privilege*, it is within 20 of 27.9. So the confidence intervals (CI)s, for a given level of confidence, are

	Hits in BNC	Per mill	CI per mill	CI per 84 mill
we	233,921	2790	2590–2990	217,660–251,160
education	23,340	278	215–341	18,060–28,644
privilege	2,338	27.9	7.9–47.9	664–4,024
granddaughter	234	2.8	0–9.1	0–764

As this makes plain, bigger numbers are far, far more dependable than smaller ones.

[19]There are exceptions for some very common words where there is a 'superheadword' for the word irrespective of part of speech. Then there are distinct headwords (printed on a grey background) for different homographs, where homographs may be distinguished on grounds of meaning, grammar or pronunciation. But these items are few and far between. A sample of 10% of the book was checked and just seven were found.

[20]Where the S- and W- scores differed in LDOCE, the one that was the best match to the COBUILD score was taken. The best match was also taken between a set of LDOCE homographs and a corresponding COBUILD headword. The two capitalised words in COBUILD occurring in the set were ignored.

Reference

Baayen, H. (1993). Statistical models for word frequency distributions: A linguistic evaluation. *Computers and the Humanities*, *26*(1–2), 347–363.

Biber, D. (1993). Using register-diversified corpora for general language studies. *Computational Linguistics*, *19*(2), 219–242.

Church, K., & Gale, W. (1995). Poisson mixtures. *Journal of Natural Language Engineering*, *1*(2), 163–190.

Church, K. W., & Mercer, R. L. (1993). Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics*, *19*(1), 1–24.

COBUILD (1995). *The Collins COBUILD English Language Dictionary. 2nd Edition.* Edited by John McH. Sinclair *et al.* London.

Crowdy, S. (1993). Spoken corpus design. *Literary and Linguistic Computing*, *8*, 259–265.

Crowdy, S. (1994). Spoken corpus transcription. *Literary and Linguistic Computing*, *9*, 25–28.

Garside, R. (1987). The CLAWS word-tagging system. In Garside, R., Leech, G. N., & Sampson, G. (Eds.), *The Computational Analysis of English*. Longman, Harlow.

Jackendoff, R. (1975). Morphological and semantic regularities in the lexicon. *Language*, *51*, 639–671.

Kilgarriff, A. (1994). The myth of completeness and some problems with consistency. In *Euralex 1994 Proceedings*, pp. 101–106 Amsterdam.

Kilgarriff, A., & Salkie, R. (1996). Corpus similarity and homogeneity via word frequency. In *EURALEX Proceedings* Göteborg, Sweden.

Kipf, G. K. (1935). *The Psychobiology of Language*. Houghton Mifflin, Boston.

LDOCE (1995). *Longman Dictionary of Contemporary English, 3rd Edition.* Edited by Della Summers. Harlow.

Leech, G., Garside, R., & Bryant, M. (1994). The large-scale grammatical tagging of text: experience with the British National Corpus. In Oostdijk, N., & de Haan, P. (Eds.), *Corpus-Based Research into Language*, pp. 47–63. Rodopi, Amsterdam.

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

Summers, D. (1992). Longman/Lancaster English Language Corpus — criteria and design. *International Journal of Lexicography*, *6*(3), 181–208.