Towards A Statistical Dictionary of Modern English: Some Preliminary Reflections

Patrick Hanks

It is sometimes claimed that certain types of dictionary do, or should, arrange the senses of each word in order of current frequency. No dictionary exists that actually does this, at any rate in English. The various (unpublished) attempts that I have been involved in over the past two decades to design such a dictionary have served merely to illustrate how important the order of senses is to dictionary discourse. Both the construction and the interpretation of sense 2 of any given word is influenced by what has been said about sense 1. A dictionary entry arranged in strict order of frequency comes across as a confusing jumble, chopping and changing madly from one theme to another and then back again. Grouping related senses together is a necessary component of clear explanation. It should be easy enough, then, one might think, to START with the most common meaning. But even here there are problems. The most common use of a word may in fact be a conventional metaphor rather than a literal meaning. For example, there is good evidence that the word 'torrent' is used far more often to refer to outpourings of speech or ideas than to mountain streams. Which of these uses should be explained first? Those who believe in notions such as 'literal meaning' or 'core meaning' (howsoever defined) may feel that it is simplistic to place a conventionalized metaphor as the first definition, and that to do so yields a description of the contemporary language every bit as distorted as that yielded by arrangement of senses on historical principles.

Nowadays, large computerized corpora of texts are becoming available to lexicographers as sources of evidence. The notion explored in this paper is that dictionaries should give explicit statistical information about the words and senses observable in the evidence. The benefits to the dictionary user are legion, and should be obvious. Just four examples will suffice here:

1 The confusion between 'most important meaning' and 'most common meaning' could be resolved. The first sense listed would be the one judged by the lexicographers to be most important (by whatever criteria of importance they wish to use), not necessarily the most frequent.

2 The construction of regular forms by grammatical extrapolation would be identified for what it is. For example, the entry for 'emblazon' might show the following statistics:

emblazon	2
emblazons	0
emblazoning	0
emblazoned	10

The forms 'emblazons' and 'emblazoning' would be identified as more potential than actual.

3 Those whose aim is to write as simply and clearly as possible would be able to use the dictionary to check that they had used only the most common meanings of common words.

4 The dictionary could be used by computer scientists, AIers, and others who wish to create a 'core vocabulary' into which all other uses can be decom posed.

***

How, then, would a statistical dictionary of modern English be constructed, and what problems lie ahead?

The pioneering work on word frequency in English is that of Nelson Francis and Henry Kučera (1982) at Brown University. The Brown Corpus consists of one million words of written American English from the year 1961. The texts were carefully selected to represent a number of different genres — in all, five hundred categories or genres are claimed. In the twenty years following the collection of the corpus, it was *tagged* — that is, each occurrence of each word received one of a set of 87 tags, so that, for example, the infinitive marker 'to' could be distinguished from the preposition 'to'. It was also *lemmatized* — that is, the various forms of each lexical item were identified so that they could be grouped together for certain analytical purposes; for example, 'go', 'goes', 'going', 'gone', and 'went' are all recognized as part of the lemma GO.

For lexicographic purposes, the Brown Corpus has a number of shortcomings. In the first place, the corpus size of one million words is far too small to provide a reasonably comprehensive account even of the the common conventional uses of common conventional words, let alone the less common conventions that are the stock-in-trade of most lexicographers. It is not intuitively plausible, for example, to say that 'hoard', 'hoary', 'hobnob', and 'hod' are not common conventional words in English, but these words, as it happens, do not occur in the Brown Corpus. All of them are in the excellent AMERICAN HERITAGE DICTIONARY, which had access to (but was evidently not limited by) the Brown Corpus.

A second criticism with the Brown Corpus is that it is restricted to written, edited, published English. There is no spoken component, but of course spoken genres are as numerous, and complex, and important as written genres.

A third criticism, from the lexicographers' point of view, is that the analysis of the Brown Corpus carried out during the past twenty-five years has not yet, so far as I know, attempted to make semantic distinctions. Thus, although the analysis clearly distinguishes between 'tear' verb and 'tear' noun, it does not distinguish between the two nouns spelled 'tear', one pronounced [tɪə] and the other [tɛə]. This is clearly a major shortcoming for any lexicographer interested in the statistics of word use. It is a profoundly interesting question how this particular shortcoming might be rectified. It is clear, I hope, from what has been said already, that any such semantic analysis needs to be based on a larger corpus than one of 1m words. The question arises, how much larger?

The experience of the COBUILD DICTIONARY may shed some light on this. Very nearly every sense of every word in the COBUILD DICTIONARY is supported by corpus evidence. As a general rule, at least two pieces of independent evidence were required to satisfy to COBUILD criterion for entry. It is worth noting that the COBUILD headword list is quite a lot smaller than that of its rivals, LDOCE and OALDCE. Now, this headword list is based on the common conventional words found in a corpus of 18m words. Much of the preliminary lexical analysis was carried out on a corpus of 7.3m words. At this level of frequency, it was clear to the lexicographers that a corpus of 7.3m words is insufficient to support a serious learners' dictionary. For example, the words 'embezzle', 'kebab', 'maisonette', and 'skive' are not in the 7.3m word corpus. However, they are found in the reserve corpus of 10.6m words which had been compiled in Birmingham by the end of 1985.

With a corpus of 18 million words, the gap between expectation and fulfilment had closed to the point where lexicographers were prepared to say, if the word (or sense) does not occur in a selection of 18m words of current English, can it be the sort of word that foreign learners need to learn how to use?

Examples of words not in the COBUILD corpus (although they are found in many current dictionaries) include 'hobgoblin', 'hogfish', and 'hogshead'. Of course, these are words of considerable interest to native speaker users of dictionaries, precisely because they are so rare. A statistical dictionary should be able to distinguish between words with a current frequency of less than one in 10 million and words with a current frequency of say, less than one in a 100 million. Ideally, statistical dictionaries of the future will be able to go on to compare the *current* frequency of certain words with their frequency in surviving texts of selected periods in the past. Here, I confine myself to problems of contemporary English.

As it happens the English language can be divided into two classes of words: a very small number that are extremely frequent and a very large number that are very rare. There are surprisingly few words of moderate frequency. The single word 'the' accounts for over 6% of all English uttered today. The 10 most common words of English-'the', 'of', 'and', 'to', 'a', 'in', 'that', 'it', 'I', and 'was'-account for over 23% of modern English text. If we extend the calculation to the top 2,000 lemmas (that is approximately 5,000 word types) we find that these account for 87% of all text. But they account for only 2% of all the word types to be found in a corpus. The Birmingham corpus consists of 260,000 word types. Of these approximately 1/2 occur only once.

Thanks to the work of Francis, Kučera, Sinclair, Clear, and others these statistics are becoming well known. They are relatively easy to discuss intelligibly. However, statistics of word use are quite another matter. It is a worthy aim to give the relative frequency of each sense of a word, but a glance at a few entries in any selection of contemporary English dictionaries will show that there is virtually no agreement among the dictionary makers on how the uses of a word are to be divided into senses. What is worse, when we try to map a collection of actual uses on to the dictionary senses, we find that it is very often impossible to decide unambiguously where to assign some uses. Even worse still, some of the common patterns of use are not clearly identified by the dictionaries.

Let us look at an example. The analysis of complex words takes hours if not days, so it is hard to find an example that is complex enough to illustrate the point while simple enough to discuss in a few minutes. The adjective 'broad' will serve to show some of these points. Figure 2 shows the senses of 'broad' as given in a highly respected American dictionary. Figure 3 shows a selection of evidence from the COBUILD 7.3 million word corpus for this word and for its inflected forms 'broader' and 'broadest'. How does this data map onto the dictionary?

In the first place, it is often hard to assign many of the uses unambiguously to a particular dictionary sense. Is "a broad bustling square" (line 28) assignable to sense 1 ("of large extent from side to side; wide") or sense 2 ("having great extent or expanse; spacious")? Are "a broad conference on disarmament" (line 38) and "the broad consensus within the Labour movement" (line 38) assignable to sense 8 ("wide in range; not limited") or to sense 9 ("main or general; not detailed")? For a statistical dictionary, decision procedures will need to be developed to cope with such problems. One such procedure would surely be to create a category "unassigned" for the rag-bag of odd uses that always turn up in a corpus, but "unassigned" should, ideally, not be extended to include ambiguous uses. This probably implies that new, more sharply differentiated sets of definitions will be called for in our statistical dictionary. It will probably consist of fewer sense, each having broader scope, than is usual in most modern dictionaries.

The criteria for assignment of uses to sense will need to be developed explicitly. Clearly, the surrounding context is what enables human lexicographers to decide whether a particular use belongs with sense 1 or sense 2. Can automatic procedures be developed to enable computers to help in the proposed assignment? For example, 'Charlotte's broad back' and 'a woman with broad shoulders' plausibly belong in the same sense category of 'broad'. If 'back' and 'shoulders' both bore the semantic tag 'BODYPART' as well as the grammatical tag 'NOUN', automatic assignment might become possible. It would then be a matter for principled decision whether to rest content with the tag BODYPART on 'forehead', 'nose', 'eyes' etc. or whether to subclassify these as, say, FACEPART. This decision would lead either to lumping or splitting of senses, in precise parallel to the current differences of taste observable among monolingual lexicographers.

The tags thus used for sorting would accumulate multiply on each lexical item (type, not lemma) in the corpus, according to the judgement of the lexicographic analyst. It is important to note that tags would be assigned to word types in the context surrounding the key word, but never directly to the key word itself.

The lexical types would themselves operate as the most delicate type of tag. For example, it is noticeable that 'broad forehead' and 'broad sense' occurs three times in the Birmingham 7.3m word corpus, 'broad daylight' and 'broad subject' four times, and 'broad shoulders' eight times.

These immediate collocations deserve to be reported by a statistical dictionary if the distribution warrant it: that is, for example, if the distribution of 'shoulders' in the context of 'broad' is significantly greater than would be the case following a random scatter of the word 'shoulders' through the corpus into all slots where a plural noun is permissible.

I said above that the statistics should report the patterns of types rather than (or at any rate in addition to) lemmas. The distribution of uses of 'broad' turns out to be rather different from those of 'broader' and 'broadest'. Uses of the comparative with nouns denoting geographical features (street, avenue, river, etc.) and body parts (shoulders, back, forehead, etc.) are remarkably rare, while use with abstract nouns, in particular 'a broader *range*' are common. The superlative is dominated by the phrase 'in the broadest *sense*'. This fact is hinted at in the 7.3 million word corpus, but since there are only 6 occurrences of 'in the broadest sense' it is not clear whether these 6 occurrences really do mean that this phrase accounts for 50% of all uses of 'broadest'. A glance at the Birmingham reserve corpus confirms that this really is so: the phrase accounts for 6 out of 12 occurrences of this word in the reserve corpus.

We have only had time to touch briefly on some of the issues involved in approaching a statistical dictionary of modern English. I conclude by summarizing some of the main features which, in my view, a statistical dictionary should have, before throwing the subject open for discussion.

1. It needs to be based on a very large sample. Clearly, 18m words is only just adequate for a statistical dictionary, and need to be supported by some rather sophisticated sampling techniques on truly huge corpora, preferably infinite in size, so that the refinement of statistics could be a continuous process. This is important for the statistics of collocation and individual word forms. Rarer and rarer words get progressively less interesting in themselves.

2. Statistics for types and lemmas need to be distinguished, as indeed they are in Francis and Kučera (1982).

3. Techniques for assigning uses to senses need to be developed more explicitly than they have been to date, and in particular collocations at all grammatical levels (clause, 'group' or 'phrase', word class, semantic class, lemma, lexeme, grammatical word, 'graphic word' or 'word type'). Statistics of collocation need to be stated explicitly as well as statistics of meaning.

References

Cited Dictionaries

AMERICAN HERITAGE DICTIONARY OF THE ENGLISH LANGUAGE (AHD). 1969/82. W. Morris, M. Berube (eds.). Boston: Houghton-Mifflin.

COLLINS COBUILD ENGLISH LANGUAGE DICTIONARY (CCELD). 1987. J. Sinclair et al. (eds.). London and Glasgow: Collins

Other Literature

Francis, Nelson and Henry Kučera. 1982. *Frequency Analysis of English Usage.* Boston: Houghton-Mifflin.