4. Is There a Basic Spoken Vocabulary: Technology and Common Sense[1]

Michael McCarthy

Abstract

A computer-generated list of the 2000 most frequently used words is an invaluable starting point for organizing a methodical approach to vocabulary building, because it helps determine clear categories. This analytical work to refine the raw data of the list and observe these categories is absolutely necessary. For example the speaker who cannot use interactive words such as *actually, basically, really, pretty, quite*, or *literally* is an impoverished speaker from an interpersonal standpoint. Speakers also need high-frequency discourse markers such as *I mean, right, well, so, good, you know*, and *anyway* to mark openings and closings and many other exchanges. This article proposes nine indispensable categories that are a necessary part of a minimum two thousand word basic vocabulary.

The search for a basic or 'core' vocabulary of English is not a new undertaking. For example, West's (1953) *General Service List* is still considered by many to be unsurpassed as a condensation of the vocabulary that was (at that time) useful for day-to-day communication, and abridged or 'essential' versions of larger dictionaries for language learners have kept the debate alive ever since. For many decades, pedagogical linguists have attempted to grade vocabulary for the purposes of syllabus design, creating word-lists for different levels of proficiency or for testing, and presenting beginners' or elementary vocabulary in teaching materials, often based on the potent combination of intuition and experience. In the last decade, computerised frequency counts of lexical items in large text corpora have assumed greater importance. These include the CANCODE corpus, which I and my colleague Ronald Carter co-direct (see McCarthy 1998). We now have frequency lists based on spoken language which bring within our grasp some of the answers to the question: What vocabulary is used most frequently in day-to-day spoken interaction? CANCODE stands for 'Cambridge and Nottingham Corpus of Discourse in English' The corpus was established at the School of English Studies, University of Nottingham, UK, and is funded by Cambridge University Press, with whom the sole copyright resides. CANCODE, in its turn, forms part of the 100-million-word written and spoken Cambridge International Corpus, also copyright Cambridge University Press. The CANCODE

corpus consists of five million words of transcribed conversations. The corpus tape-recordings were made in a variety of settings across the islands of Britain and Ireland, with a wide demographic spread. For further details of the corpus, see McCarthy (1998).

In computer frequency counts, there is usually a point where frequency drops off rather sharply, from hard-working words which are of extremely high frequency to words that occur relatively infrequently. Thus the frequencies do not decline at a regular rate, but usually have a point where there is a sudden drop to low frequency. rarer words. The point where such a drop is discernible could be seen as a boundary between the 'core' and the rest, though the picture may not always be clear cut, and might be expected to vary a little from corpus to corpus. In a three-million-word sample of the CANCODE corpus, what is apparent is that, round about 1800-2000 words down in the frequency ratings, the graph begins to drop more steeply, with a marked decrease in the number of words that occur more than 100 times. We can therefore conclude that words occurring approximately 100 times or more in this sample belong to some sort of heavy-duty core vocabulary, amounting to about 1500-1600 words. The first 2000 words, where the graph seems to start upon its steeper descent, are those which occur around 75 or more times in the corpus. It is reasonable to suppose, therefore, that a round-figure pedagogical target of the first 2000 words in order of frequency will safely cover the everyday core with some margin for error.

The top 40 most frequent words occur well in excess of 10,000 times in the sample, and thus have a heavy duty application because of their frequent use. However, questions arise as to their place in a 'vocabulary' list. Very many of the words clearly belong to the traditional province of grammar/function words, in that they are devoid of lexical content. These include articles, pronouns, auxiliary verbs, demonstratives, basic conjunctions, etc. Most language teachers would consider these to be part of the grammar syllabus. The types of meaning they convey (e.g. the deictic meanings of pronouns such as *I* and *you* and the demonstratives, or the additive and adversative meanings, respectively, of conjunctions such as *and* and *but*) are considered to be grammatical rather than lexical (vocabulary) ones. Another problem with the top 40 list takes us to the question of fixed phrases, or lexicalised 'chunks' extending over more than one word. Word #14 (*know*) and word # 37 (*think*) prove to be so frequent mainly

because of their regular collocation with *you* and *I*, respectively, in the formulaic utterances *you know*, and *I think*.

The top 40 list shows that arriving at the basic vocabulary is not just a matter of instructing the computer to list the most frequent individual forms (or 'tokens'), and considerable analytical work is necessary to refine the raw data. Nonetheless, the computer-generated first 2000 word list is an invaluable starting point, for a good many reasons, not least because fairly clear categories emerge from it which offer the potential for an organised pedagogy (in the sense that few language teachers would ever propose simply working one's way down the list as a viable methodology for vocabulary building). Those categories are what I now want to go on to illustrate. If, on the basis of general professional consensus, we exclude as a category the closed-system grammar/function words, the remainder of the 2000 word list seems to fall into approximately nine types of item, all equally important, and not prioritised here.

1. Modal vocabulary

Modal items are those which refer to degree of certainty or necessity. Clearly these include the closed class of modal verbs usually taught as grammar (*can, could, may, must, will, should*, etc.), but the list contains other very high frequency modal items too, for example the verbs *look, seem* and *sound*, the adjectives *possible* and *certain* and the adverbs *maybe, definitely, probably* and *apparently*. Some of these may strike teachers as more 'intermediate' level words, and yet their frequency is so high in everyday talk that excluding them from the elementary level would need some other justification (e.g. such as avoiding duplication of close synonyms and economising on cognitive load).

2. Delexical verbs

This category includes extremely high-frequency verbs such as *do, make, take* and *get* in their collocations with nouns, prepositional phrases and particles. They are termed delexical because of their low lexical content and the fact that statements of their meaning are normally derived from the words they co-occur with (e.g. compare *to make it to a place* with *to make a mistake* or *to make dinner*). However, one problem associated with the massive frequency of the delexical verbs is the fact that their low lexical content has to be complemented by the lexical content of the words they combine with, and those collocating

words may often be of relatively low frequency (e.g. *get a degree, get involved, make an appointment*), or may be combinations with high-frequency particles generating semantically opaque phrasal verbs (e.g. *get round to doing something, take over from someone*). In the language class, the delexical verbs cannot be taught in isolation, without reference to their collocations, so the task becomes one of ascertaining the most frequent and useful collocating items from lower down in the frequency list, such as *get a job, take something back, make coffee*, etc., which might occasionally involve words from outside of the top 2000, but which are necessary to provide authentic contexts for the learning of the delexical verbs.

3. Interactive words

The core 2000 word list contains a number of items whose function is to present speakers' attitudes and stance. These are absolutely central to communicative well-being, to creating and maintaining appropriate social relations. They are therefore not a luxury, and it is hard to conceive of anything but the most sterile and banal survival-level communication occurring without their frequent use. The speaker who cannot use them is an impoverished speaker, from an interpersonal viewpoint. The words include *just, whatever, thing(s), a bit, slightly, actually, basically, really, pretty, quite, literally*. Their high frequency in speech underlines their vital role in face-to-face communication. For example, *just* occurs more than 4000 times per million words in the spoken corpus, compared with only 1400 times per million words in a 5-million word segment of the written component of the Cambridge International Corpus.

The interactive words may variously soften or make indirect potentially face-threatening utterances, purposively render vague or fuzzy acts of lexical categorisation in the conversation, or intensify and emphasise affective stance towards the content of utterances.

4. Discourse markers

The core spoken vocabulary contains high-frequency discourse markers that function to organise the talk and monitor its progress. The most common ones occurring in the top 2000 include *I mean, right, well, so, good, you know, anyway*. Their functions include marking openings and closings, returns to diverted or interrupted talk, topic boundaries and exchange completions. There is evidence

to suggest that native speakers are poor judges of the all-pervasiveness of such markers in their own talk (Watts 1989), and indeed their frequent use may be perceived by language purists to be a sign of bad or sloppy usage, and yet all the evidence in the spoken corpus is that the markers are ubiquitous in the conversation of educated native speakers. The high-frequency discourse markers also have little lexical content in the conventional sense of the word, and present a problem to language pedagogy, which has traditionally divided teaching into grammar teaching and vocabulary teaching with items such as discourse markers not fitting happily into either. In short, there is no ready-made pedagogy for this category of items, a point we shall return to in the concluding section.

5. Basic nominal concepts

Into this category fit a wide range of nouns of very general, non-concrete and concrete meanings, such as *person, problem, life, noise, situation, sort, trouble, family, kids, room, car, school, door, water, house, TV, ticket*, along with the names of days, months, colours, body-parts, kinship terms, other general time and place nouns such as the names of the four seasons, the points of the compass, and nouns denoting basic activities and events such as *trip* and *breakfast*. Additionally, one may include here semi-grammatical items such as *both, something, everything, sometimes*.

However, interesting problems arise in terms of the closed-set nature of some of these nouns. In any corpus, items apparently belonging to closed sets will not necessarily occur with equal frequency. There is a wide discrepancy, for example, in how the seven days of the week occur, with the weekend days, Friday and Saturday, achieving almost double the frequency of 'low' days such as Tuesday and Wednesday. There may well be cultural reasons for such unequal distribution (in Westernised, Christian societies, Monday is considered the start of the working week; Friday and Saturday are associated with the week's end and leisure, etc.), and the corpus can indeed be used as a cultural 'window' for language teaching purposes, but for the goal of imparting a basic vocabulary of communication, only the most purist of corpus-adherents would propose a pedagogy wherein the elementary level would only teach five of the seven weekday names, leaving the low frequency Tuesday and Wednesday till the intermediate level. Thus corpus statistics need to be combined with a notion of psycholinguistic usefulness and

the availability (*disponibility*) of items in the native speaker lexicon. Pedagogical decisions may override these awkward but fascinating statistics, and most teachers will agree that it makes good sense to teach basic closed sets as completely as is practically possible. However, some closed sets are very large (e.g. all the possible body parts, or the names of all countries in the world), and in such cases, the frequency list is very helpful for establishing priorities.

6. General deictics

Deictic items relate the speaker to the world in relative terms of time and space. The most obvious examples of deixis are words such as the demonstratives, where *this box* for the speaker may be *that box* for a remotely placed listener, or the speaker's *here* might be *here* or *there* for the listener, depending on where each participant is relative to the other. The corpus, in addition to the demonstratives and *here* and *there*, contains key items with relative meanings such as *now, then, ago, away, front, side* and the extremely frequent *back* (in the sense of opposite of front, but mostly in the sense of returned from another place). *Back* occurs 3722 times, most frequently in the clusters *go/come/get back, the back of* (something), *at/in/on the back, put/take* (something) *back*, and is clearly a core word in spoken English. Similarly *being away* and *being out* are of very high frequency and distinguish two different everyday deictic concepts.

7. Basic adjectives

In this class there appear a number of adjectives for communicating everyday positive and negative evaluations of people, situations, events and things. These include *lovely, nice, different, good, bad, horrible, terrible*. Questions of usefulness and near synonymy are raised, and close observation of actual occurrences in the corpus, and ascertaining how the different adjectives enter into lexico-grammatical patterns is vital for resolving the issue of what to include, what may be delayed till later stages in the vocabulary teaching and learning operation, etc. *Horrible* and *terrible*, for example, although close in meaning, and although almost identical in frequency, seem to have a preference for patterning with nouns denoting people, things or situations (in the case of *horrible*) and situations but not people (in the case of *terrible*). These are broad preferences, and can only be stated in probabilistic rather than absolute terms, but they can prove significant in the decision to include both words in a vocabulary syllabus, even though their meanings may seem to overlap (see McCarthy and O'Dell 1999:48).

8. Basic adverbs

Many adverbs are of extremely high frequency, especially those referring to time, such as *today, yesterday, tomorrow, eventually, finally*, frequency and habituality, such as *usually, normally, generally*, and manner and degree such as *quickly* (but not *slowly*), *suddenly, fast, totally, especially*. Also extremely frequent are sentence adverbs such as *basically, hopefully, personally and literally*, which function to evaluate utterances and which reflect speaker stance. This class of word is fairly straightforward, but it should be borne in mind that some prepositional phrase adverbials are also extremely frequent, such as *in the end*, and *at the moment*, which occur 205 and 626 times, respectively. The raw frequency list hides the frequency of phrasal combinations, and extra research is needed to ensure that the most frequent phrasal items are not lost from the basic vocabulary.

9. Basic verbs for actions and events

Beyond the group of delexical verbs, there are, of course, a number of verbs denoting everyday activity, such as *sit, give, say, leave, stop, help, feel, put, listen, explain, love, eat, enjoy*. It is worth noting that the distribution of particular tense/aspect forms may be relevant in considering priorities in the basic vocabulary. Of the 14,682 occurrences of the forms of the verb *say* (i.e. *say, says, saying, said*), 5,416 of these (36.8%) are the past form *said*, owing to the high frequency of speech reports in the spoken corpus With *tell*, this is also true: almost 30% of all examples are past tense *told*. With *give*, the picture is much more even: the simple past form, *gave*, accounts for only 15% of all occurrences of the verb. Such differences may be important in elementary level pedagogy, where vocabulary growth might outstrip grammatical knowledge, and a past form such as *said* might be introduced to frame speech reports even though familiarity with the past tense in general may be low or absent on the part of the learner.

Conclusion

The ability to generate word lists based on frequency is one of the most useful tasks a computer can perform in relation to a corpus, and especially with spoken data, where a clear core vocabulary based around the 1500-2000 most frequent words seems to emerge, a vocabulary that does very hard work in day-to-day communication. However, we have seen that raw lists of items need careful

evaluation and further observations of the corpus itself before a vocabulary syllabus can be established for the elementary level. Armed with the complex information a computerised list can give, the teacher, syllabus designer or materials writer can elaborate a more use-centred vocabulary pedagogy at the elementary level and provide useful and usable language items even to very low level learners. Until recently, word lists were derived from intuition or from written text sources; our ability nowadays to produce lists based on spoken data considerably enhances our potential for teaching the spoken language more effectively and authentically.

Note

1 A longer, more statistically detailed version of this article may be found in *SELL*, Issue No.2, 1999. University of Valencia Press, Valencia, Spain, Department of English and German Philology.

Michael McCarthy is Professor of Applied Linguistics at the University of Nottingham. He is author of *Language as Discourse* (Longman, 1994, with Ronald Carter), *Exploring Spoken English* (Cambridge University Press, 1997, with Ronald Carter), *Spoken Language and Applied Linguistics* (Cambridge University Press, 1998) and *English Vocabulary in Use*, Upper Intermediate and Elementary (Cambridge University Press, 1994 and 1999, with Felicity O'Dell). From 1994 to 1998 he was co-editor of *Applied Linguistics* (Oxford University Press). He is co-director (with Ronald Carter) of the 5-million word CANCODE spoken English corpus project, sponsored by Cambridge University Press, at the University of Nottingham. He lives in Cambridgeshire, UK, where he writes books, grows fruit and vegetables and plays Irish traditional music on the fiddle.

References:

McCarthy M J (1998) *Spoken Language and Applied Linguistics*. Cambridge: Cambridge University Press

McCarthy M J and O'Dell F (1999) *English Vocabulaty in Use. Elementary*. Cambridge: Cambridge University Press

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

Watts R J (1989) Taking the pitcher to the 'well': native speakers' perception of their use of discourse markers in conversation. *Journal of Pragmatics* 13: 203-37