

What constitutes a basic vocabulary for spoken communication?

Published in *Studies in English Language and Literature*, (1999) vol.1: 233-249.

Michael McCarthy

University of Nottingham

Great Britain

1 Introduction

The quest for a basic or ‘core’ vocabulary of English is not a new one. Ogden’s (1930) attempt to formulate an 850-word lexicon that would have as wide a communicative range as possible using a minimum number of words of general meaning, is a famous one (for a description and comments, see also Carter and McCarthy 1988: 2-6). West’s (1953) *General Service List* is still considered by many to be unsurpassed in its usefulness, and other word lists that distinguish between core or elementary level words and more advanced words (e.g. Hindmarsh 1980), and cut-down or ‘essential’ versions of larger dictionaries for language learners have kept the quest alive. On the theoretical level, Carter (1987) has discussed the ‘core’ properties of certain words and suggested criteria for any word to qualify as a core word. And for many decades, pedagogical linguists have attempted to grade vocabulary for the purposes of syllabus design, creating word-lists as targets for different levels of accomplishment or for testing purposes, and presenting beginners’ or elementary vocabulary in teaching materials, often simply based on a powerful admixture of intuition and experience.

In the last fifteen years or so, the awesome power of computational analysis has come along to assist the enterprise, first off mostly in the form of frequency counts of lexical items in written texts, but latterly supplemented by work springing from the ever-increasing number of spoken corpora . These include the London-Lund

corpus (see Svartvik and Quirk 1980), the CANCODE corpus, on which the present paper is based (see McCarthy 1998), the spoken element of the British National Corpus (see Rundell 1995), the Santa Barbara corpus of spoken American English (see Chafe *et al* 1991), and the University of Valencia corpus of spoken Spanish (see Briz 1995). It is now possible to have access to computer-generated frequency lists based on spoken language which bring within our grasp some (but not all) of the answers to the question: what vocabulary is used most frequently in day-to-day spoken interaction? The present paper explores this question and offers a tentative sketch of a basic spoken vocabulary, while also rehearsing some of the issues and problems that arise.

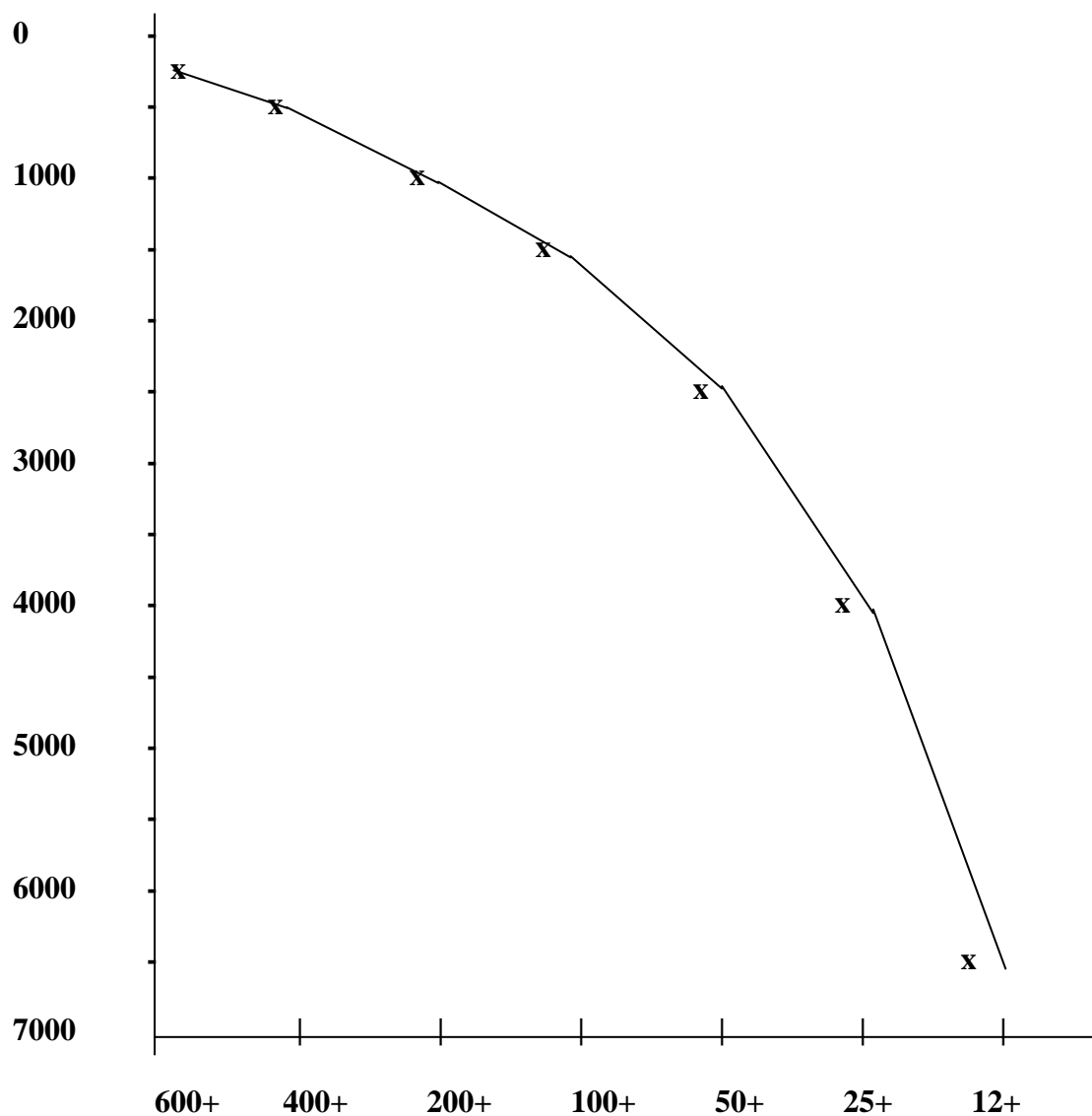
The present paper is based on a 3-million-word sample of the 5-million-word CANCODE spoken corpus. CANCODE stands for >Cambridge and Nottingham Corpus of Discourse in English=. The corpus was established at the Department of English Studies, University of Nottingham, UK, and is funded by Cambridge University 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 including private homes, shops, offices and other public places, and educational institutions (though in non-formal settings) across the islands of Britain and Ireland, with a wide demographic spread. For further details of the corpus and its construction, see McCarthy (1998).

2 How big does a basic spoken vocabulary need to be?

There is no easy answer to the question posed in this section, except to say that, in computer-based 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, in other words, the frequencies do not

decline at a regular rate, but usually have a point where there is a sudden drop to low frequency. The point at 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. Figure 1 shows how frequency drops off in the CANCODE sample upon which the present paper is based. The horizontal axis of the graph shows frequency of occurrence (i.e. 12+ indicates words occurring more than 12 times in the sample, 400+ words occurring more than 400 times, etc.). The vertical axis shows how many words in the corpus actually occur at the given frequencies (for example, around 2500 words occur fifty or more times in the sample).

Figure 1: Decline in frequency of occurrence of words in the CANCODE sample



occurrences

What is apparent is that, round about 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 (this occurs most noticeably from word number 1500 onwards in the list). 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 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. It is to these 2000 words that we now turn, for they provide us not only with raw data, but with an insight into the types of words that the core vocabulary embraces. The types are as important as the words themselves, since they tell us much about the elements that compose interaction from a lexical viewpoint.

3 The most frequent words

Figure 2 shows the 40 most frequent items in the 3-million-word sample, as identified by a computer count.

Figure 2: 'Top 40' words of the CANCODE sample.

N	Word	19	IS
1	THE	20	SO
2	I	21	LAUGHS
3	AND	22	WE
4	YOU	23	ON
5	TO	24	HAVE
6	A	25	WELL
7	IT	26	NO
8	YEAH	27	DO
9	THAT	28	WHAT
10	OF	29	FOR
11	IN	30	LIKE
12	MM	31	RIGHT
13	WAS	32	OH
14	KNOW	33	JUST
15	ER	34	HE
16	IT'S	35	BE
17	THEY	36	ALL
18	BUT	37	THINK

38	THIS
39	THERE

40	GOT
----	-----

All of these words occur well in excess of 10,000 times in the sample, and thus perform heavy duty in terms 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 stock-in-trade of the grammar syllabus, and in most beginner and elementary level language learning materials they get careful attention and are taught as closed systems within the grammar. 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 functions rather than lexical ones.

This still leaves a good number of words unaccounted for, and, indeed, a number of items whose entitlement to the label *word* is open to debate. The computer has, for example, noted that the transcription of laughter in the form of the word *laughs* (word # 21) occurs with massive frequency. This can be factored out by tweaking the software, but a sociolinguist or conversation analyst might observe that laughter performs important responsive functions and may signal boundaries in the discourse, etc, in the same way that *mm* (word # 12) does, but *mm* may well be considered a more worthy candidate for the title of word or vocabulary item on the grounds that it expresses meanings such as acknowledgement, topic pausing, agreement, hesitation, etc., and does seem to have subtle sub-variants which careful analysis can reveal as distinguishing very finely between types of function (see Gardner 1998). What is more, *mm* may be realised quite differently in different languages/cultures, and it may indeed be a useful vocalisation to learn, even if we deny it full status as an item of vocabulary. Put another way, there are arguments for suggesting that a vocabulary list, defined as a list of non-grammatical meaning-

resources, is not necessarily co-terminous with a word list, especially in discourse-based approaches to language description and pedagogy. We may here note that word # 15 (*er*) and word # 32 (*oh*) fall into a similar category to *mm*.

Another problem raised by the top 40 list takes us in the opposite direction, away from the consideration of minimal vocalisations, into 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*. To use frequency counts properly so as to yield up the useful information they often hide, one must also carry out cross-comparisons as to potential co-occurrence of words in the list. For example, breaking down word # 39, *there*, into its occurrences as a spatio-temporal deictic and its co-occurrences with *is/are* as an existential subject in a formulaic configuration may lead to some redistribution of the items within the list.

A further problem arises with inflexions of base-forms. Word # 40, *got*, may prompt us to search for its allied forms *get(s)/getting*, and their lower position in the list of 2000 (*getting* occurs a mere 1500 times in our sample, compared with just under 12000 occurrences of *got*) may not necessarily deny the lemma *GET* (the form that represents the base *and* all its inflexions) a very high place in the basic vocabulary list, and there is every good reason for adding together the frequencies of allied forms. For this reason, linguists such as Nation (1997) prefer to talk of word *families* when considering questions of vocabulary size.

All in all, then, 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 forms, 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 (insomuch as 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 the main body of this paper is devoted to illustrating. If, on the basis of general professional consensus, we exclude as a category the closed-system grammar/function words (although we shall return to re-consider them at the end of this paper, the remainder of the 2000 word list seems to fall into approximately nine types of item, which I shall examine in turn. They are not presented in any prioritised order, and all may be considered equally important as components of basic communication.

4 The nine broad categories of a basic spoken vocabulary

4.1 Modal items

Modal items are those which carry meanings referring to degree of certainty (epistemic modality) or necessity (deontic modality). Clearly the best candidates for such meanings in the 2000 word list are the closed class of modal verbs (*can, could, may, must, will, should, etc.*), but the list contains other very high frequency items that carry related meanings. These include lexical modals such as 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). Typical examples from the data are:

(1)

- A: I mean how long does it take to get down? It's four hours isn't it? What about if we say we'll get there for twelve?
- B: That **sounds** all right to me.

(2)

[Message on a telephone answering machine]

Caller: It's to tell you that erm, I'm, it's Friday at ten o'clock, and I'm leaving the hotel now and going to the garage to pick up the car. It's all ready **apparently**.

In each case the degree of certainty is modified in the same way that closed-class modal verbs such as *could* and *must* are characteristically used. To suggest that the domain of modality be expanded beyond the closed-class modal verbs is not a new idea; several linguists have advocated this based, on the frequent occurrence in written texts of a wider range of modal items (Holmes 1988) or on sociolinguistic 'fieldwork' (Stubbs 1986). The spoken corpus statistics underscore this earlier work and provide compelling evidence of the ubiquity of modal items in everyday communication.

4.2 Delexical verbs

This category embraces 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*). In the case of *do* and *get* a distinction has to be made between their auxiliary-verb functions: *do* in emphatic, negative and interrogative verb phrases and in tags, and *get* in the *have got* (possessive), *have got to* (modal) and *get*-passive constructions, the last being far more frequent in spoken data than in written (Carter and McCarthy 1999). Typical contexts are:

(3) When I **got** to work I thought I'd see you at work.

(4) I hope I **get** some money for it.

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 language pedagogy, 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.

4.3 Interactive words

The core 2000 word list contains a number of items whose function is to represent speakers' attitudes and stance towards the content communicated. 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. Some examples follow:

(5) You fly from Birmingham to Berlin, and then get a taxi **or whatever**, from the airport to the railway station.

(6) I'm **just** ringing up to enquire whether there was any more definite news.

(7) So it's **a bit** worrying **really**.

4.6 Discourse markers

The core spoken vocabulary contains high-frequency discourse markers whose function is to organise the talk and monitor its progress. A range of such items has been recognised by linguists such as Schiffrin (1987) and Fraser (1990), and 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. They are, therefore, like the interactive words dealt with in section 4.3, an important feature of the non-propositional elements in any discourse, and, for conversational participants they provide a resource for exercising control; they have an empowering function, the absence of which in the talk of any individual conversational participant leaves him/her potentially disempowered and at risk of becoming a second-class participant.

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. Examples include:

(8) **Anyway**, I'll have to go because I've got to ring Simon.

(9)

Customer: Can I post it as fast post?

Postal assistant: **Well** if you want to send it Swift Air it costs you another two pounds seventy.

Customer: Two pounds seventy.

Postal assistant: Yeah.

Customer: But how long?

Postal assistant: **Well** it doesn't tell you.

(9) Erm, **so, you know**, next time I call, if I could get that from you it would save me ringing up Lorna and Bill and having to talk to Bill, probably.

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.

4.5 Basic nouns

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*.

These nouns, because of their general meanings, have wide communicative coverage. *Trip*, for example, can clearly substitute for *voyage*, *flight*, *drive*, and so on. 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. Figure 3, for example, shows the occurrence per million words of the names of the seven days of the week.

Figure 3: the seven days of the week in the CANCODE sample (occurrences per one million words)

Day	per 1m
Monday	98
Tuesday	61
Wednesday	57
Thursday	68
Friday	108
Saturday	102
Sunday	83

There is a wide discrepancy here, 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 (*disponibilité*) of items in the native speaker lexicon. Figure 4 shows the frequency distribution of basic colour names in the sample, where considerable variation exists, with *brown* just managing to fall inside the 75+ occurrence range considered crucial for the present paper, and *grey* falling outside of the basic vocabulary limit. *Black* occurs almost seven times more frequently than *grey*, while *red* and *green* appear to be almost identical in their values. Decisions whether to include brown and exclude grey from the basic list must be made in light of pedagogical usefulness, and most vocabulary programmes would probably wish to include both, on the grounds that *grey* is useful for describing hair colour, the sky, clothes, buildings, etc. *Purple*, on the other hand, occurring only 20 times in the sample, might indeed be considered a non-basic word, both in terms of its frequency in the corpus and its usefulness for speaking about the day-to-day world.

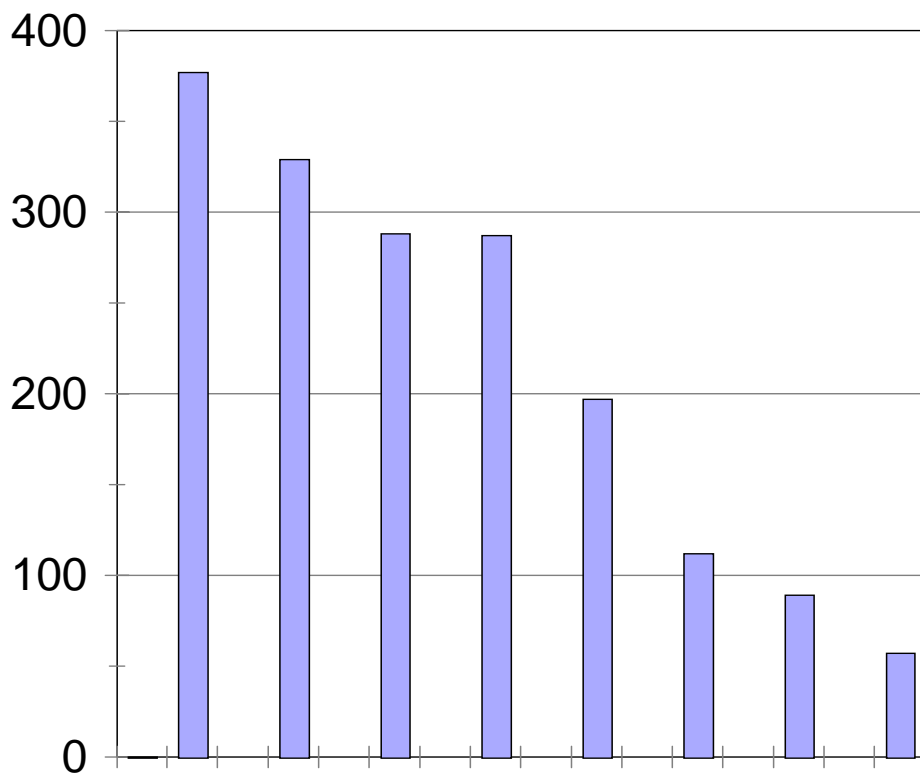


Figure 4: Occurrences of colour terms in the CANCODE sample (3m words)

Black white red green blue yellow brown grey

In terms of further examples, amongst the human body parts, *arm*, *finger*, *head*, *foot*, *nose*, *eye*, and *leg* all make it into the top 2000 list, but *shoulder*, *knee* and *wrist* do not. In the names of the four seasons, *summer* is three times more frequent than *winter*, and four times more frequent than *spring*, with *autumn* trailing behind at ten times less frequent than *summer* and outside of the top 2000 list. In the names of countries, *Italy*, *France*, *America* and *Australia* make the top 2000 list, *Spain* is right on the borderline, while *China*, *Japan* and *Canada* fall outside of the list. Once again, 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.

Another interesting issue arises with regard to the occurrence of near-synonyms within the list. For example, *trip* occurs 115 times, and thus qualifies for membership of the basic list; *journey*, its near synonym occurs only 54 times, and falls outside of the desired 75+ range. The decision would seem to be straightforward, that *trip* is the useful, frequent word, and might be able to ‘cover’ for *journey* in basic, survival-level communication. In other cases, the difference is even greater: the various verb-forms of *begin*, for example, occur 164 times; the verb forms of *start* occur 1985 times, 12 times greater frequency. Both verbs fall within the basic list, but *start* is clearly a much higher priority word. Some cases, however, are not so clear-cut: *centre* occurs 291 times, and its apparently synonymous *middle* 312 times, both very high figures and extremely close to one another. In some cases they will be interchangeable (*the middle of the village/the centre of the village* might both be equally communicative), but in others not (*the city centre/*the city middle*), and the decision might be to retain both in the list because of their different collocations.

4.6 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 each 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. Typical examples include:

- (10) She's **out** about twelve hours a day. (not at home)
- (11) I was **away** yesterday. (not at home and in a different town/region)
- (12) So you're **back** in Oxford now, right?
- (13) That would be down the **bottom right hand side**.

Deixis is also encoded in the basic verbs *go* and *come*, and *take* and *bring* (see below), and is a core function reflected widely in the 2000 word basic list.

4.7 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, different*. Once again, 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 lexicogrammatical 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 (247 occurrences and 252 occurrences in our sample, respectively) 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*) respectively. These are broad preferences, and can only be stated in probabilistic rather than absolute terms, but nonetheless such patterns of preference are evident, and 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). In other cases, degrees of intensity are involved (e.g. the mid-range *nice* compared with the stronger *lovely*) and it may be advisable to include more

than one term for the sake of interpersonal variation, enabling the user to avoid projecting a rather one-dimensional self-image. Examples include:

- (14) It was so **terrible**, the wet weather.
- (15) They're awful, **horrible** people.
- (16) What a **horrible** place, London.
- (17) It was a very **nice** dress.
- (18) Her mum was a **lovely** person.

Two other issues arise with basic adjectives. The first is a strong preference in conversational grammar for predicative position (i.e. after a verb such as *be*, *look*, *sound*, etc., rather than in attributive position before a noun). Some examples follow:

- (19) A: Was she **nice**?
- B: She was **lovely**.
- (20) They say everybody's **different**.

McCarthy and Carter (in press) note, with regard to the preference in spoken grammar for predicative adjectival positioning, that in a sample of more than 1300 occurrences of the noun *house*, it is extremely rare to find more than one adjective preceding the noun, and where further descriptive information is accumulated it is more typically effected in the post-noun-head position, as in example (21):

- (21) It's a **large** house, **lovely**, **just right**.

McCarthy and Carter compare this with many written corpus examples of *house* which have more complex pre-head modification (e.g. *a big, dirty, communal house*). The point of such observations is that it would seem that exemplification of the basic adjectives for the teaching of spoken vocabulary might be more authentic if done in predicative- position utterances, and not just in attributive position.

The second issue relating to basic adjectives is their frequent occurrence as response tokens. *Great* occurs very frequently in this function:

(22) A: I'll get back to you in the next ten minutes.

B: **Great.**

A: All right?

B: Thank you.

(23) A: I'll get them to give you a ring when they get back, okay?

B: **Fine.**

These important tokens of listenership make the difference between a respondent who repeatedly acknowledges incoming talk with an impoverished range of vocalisations or the constant use of *yes* and/or *no* and one who sounds engaged, interested and interesting. The basic adjectives do more, therefore, than just provide a *descriptive* apparatus; they offer the speaker a range of responding functions, and can be used very simply, even at elementary levels of competence, as single-word response tokens. All of these observations are part and parcel of viewing the basic 2000 word list as a communicative resource for everyday conversation rather than just as a means of representing the world at the propositional level. Indeed, one might well conclude that for many of these words, where their response function at least equals and often outweighs their descriptive function (descriptive in terms of the traditional notion of adjective as being an item describing a nominal item, either attributively or predicatively) in terms of frequency, the label *adjective* seems not entirely appropriate, since they evaluate a situation or a whole utterance, and are operating at the level of discourse rather than within the phrase or clause. The same applies to adverbs that occur with high frequency as response tokens, such as *absolutely* and *definitely*, suggesting that a contextually determined word-class with the label *response token* or *feedback token* might be more useful as a category for pedagogy.

4.9 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.

4.9 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.

5. Conclusion

The ability to generate word lists based on frequency of occurrence 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. Not least of the problems is that of widely differing frequencies for sets of items that seem, intuitively, to belong to useful families for pedagogical purposes. Equally, the list needs to take account of collocations and phrasal items, as we saw in the case of the delexical verbs, the discourse markers and the basic adverbs. But the list can also be very useful in suggesting priorities for apparent synonyms and in establishing graded information for closed sets consisting of very large numbers of items (e.g. the human body parts). 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.

References

- Briz A (ed) 1995 *La conversación coloquial: Materiales para su estudio*. Valencia: Universidad de Valencia, Facultad de Filología
- Carter R A 1987 Is there a core vocabulary? Some implications for language teaching. *Applied Linguistics* 8 (2): 178-93
- Carter R A and McCarthy M J 1988 *Vocabulary and Language Teaching*. London: Longman
- Carter R A and McCarthy M J 1999 The English *get*-passive in spoken discourse: description and implications for an interpersonal grammar. *English Language and Linguistics* 3 (1) 41-58
- Chafe W, Du Bois J and Thompson S 1991 Towards a new corpus of spoken American English. In Aijmer K and Altenberg B (eds) *English Corpus Linguistics*. London: Longman, 64-82
- Fraser B 1990 An approach to discourse markers. *Journal of Pragmatics* 14: 383-95
- Gardner R 1998 *Between speaking and listening: the vocalisation of understandings*. *Applied Linguistics* 19 (2): 204-224
- Hindmarsh R 1980 *Cambridge English Lexicon*. Cambridge: Cambridge University Press
- Holmes J 1988 Doubt and certainty in ESL textbooks. *Applied Linguistics* 9 (1): 21-44
- McCarthy M J 1998 *Spoken Language and Applied Linguistics*. Cambridge: Cambridge University Press
- McCarthy M J and O'Dell F 1999 *English Vocabulary in Use. Elementary*. Cambridge: Cambridge University Press

- Nation P and Waring R 1997 Vocabulary size, text coverage and word lists. In Schmitt N and McCarthy M J *Vocabulary: Description, Acquisition and Pedagogy*. Cambridge: Cambridge University Press, 6-19
- Ogden C K 1930 *Basic English: A General Introduction*. London: Kegan Paul, Trench and Trubner
- Rundell M 1995a The BNC: A spoken corpus. *Modern English Teacher* 4(2): 13-15
- Schiffrin D 1987 *Discourse Markers*. Cambridge: Cambridge University Press
- Stubbs M 1986 'A matter of prolonged fieldwork': notes towards a modal grammar of English. *Applied Linguistics* 7 (1): 1-25
- Svartvik J and Quirk R 1980 *A Corpus of English Conversation*. Lund: Liberläromedel
- 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