Frequency in Learners' Dictionaries

Paul Bogaards
Leiden University

The learners' dictionaries that exist for English all contain a restricted number of items. The vocabulary that is described in these dictionaries is selected on the basis of frequency of appearance in English. A far more limited number of items are marked as the most important ones, as these that all students should know at some time, because they constitute the lexical core of the language. The marking of high frequency is done in different ways in the five learners' dictionaries. The data provided are not always very useful and are sometimes inconsistent from dictionary to dictionary. An analysis is made of some samples taken from the five learners' dictionaries of English and the relevance of different types of frequency information is discussed.

1. Introduction

The learners' dictionaries that exist for English all contain restricted numbers of items. The vocabulary that is described in these dictionaries is selected on the basis of the frequency of appearance of words in English texts. The numbers of lexical units[1], i.e. (monosemous) words, word senses, and multi-word expressions, that are defined and illustrated vary between some 70,000 (for Cobuild and CIDE, i.e the first edition of what is hereafter called CALD2) and more than 90,000 (for MEDAL) or 100,000 (for LDOCE, cf. Bogaards 1996, Bogaards 2003). As these numbers are too big to be learned or known by the students for whom these dictionaries are designed, a far more limited number of items are marked as the most important ones. These are the ones that all students should know at some time, because they constitute the lexical core of the language or because they may be used to define the other words contained in the dictionary.

Dictionary	**Corpus**	**Written (N words)**	**Spoken (N words)**	**Total N words**
LDOCE4 (2003)	Longman Corpus Network	yes	yes	300 million
CALD2 (2005)	Cambridge International Corpus	yes	yes	600 million
OALD7 (2005)	British National Corpus Oxford Corpus Collection			?
Cobuild5 (2006)	Bank of English	605 million	40 million	645 million
MEDAL2 (2007)	World English Corpus			200 million

Table 1. Learners' dictionaries and their underlying corpora.

The text corpora that are used for these dictionaries contain several hundreds of millions of words, taken from a variety of sources, including spoken texts. Table 1 gives some details of these corpora. As this table shows, the numbers of occurrences (words) vary widely, from about 200 million to more than 600 million. In most cases the proportion of spoken data is not indicated; the word "yes" in Table 1 means that dictionary concerned states that written as well as spoken data are included in the corpus but that no details are given. Only Cobuild5 (p. x) mentions that about 6% of the corpus concerns 'informal spoken language' which is represented by "recordings of everyday casual conversation, meetings, interviews, and discussions as well as by transcriptions of radio and TV programmes". Be

[1] The term "lexical unit" is taken here in the sense of Cruse (1986) and covers all senses and uses that have their own definition in the dictionary.

that as it may, the dictionaries are mainly based on written data, which indeed corresponds well to the first use to which they are put: reading.

One would say that in spite of their differences in volume and in composition, these corpora are big enough to establish in a stable way the frequency of the most common elements of the language. This, however, turns out not to be always the case. Several reasons can account for this situation.

2. Frequency data in the "big five"

In the first place, although frequency certainly has played an important role in establishing the lists of core items, it is telling that most of these lists are not presented as being mere frequency lists. They are said to contain "important words" (OALD7), "common words" (MEDAL2), or are presented as "defining vocabulary" (LDOCE4). Only Cobuild5 uses the term "frequency bands". In several cases it is explicitly stated that frequency was not the only factor in choosing the core vocabulary. OALD7 (p. viii), for instance, states that the Oxford list is "not simply a list of the 3,000 most frequent words of English. It includes words which would fall outside such a list, but which are useful for learning as well as relevant for defining purposes. It is […] usefulness that is the deciding factor." CALD2 (p. vii) specifies that the "essential" words are "either extremely common […], or […] express core concepts (e.g. *asleep*)." Although in all cases frequency was one of the important aspects, if not the most important one, it is to be expected that the lists will vary from dictionary to dictionary.

Furthermore, different policies are followed to mark these important or high frequency words in the dictionaries. OALD7 only marks, with a key, the forms belonging to the "Oxford 3,000 list of important words." MEDAL2 employs a system of red stars where one star indicates a "fairly common" word whereas three stars appear with words that belong to the "most basic words of English". These stars are given not just with forms, as is done in OALD7, but with elements of a given word class. In MEDAL2 the form *bank*, for instance, has three stars for its uses as a noun (all senses taken together), but only one when it is used as a verb. Cobuild5 has a similar system of giving information about frequency, but uses diamonds as symbols. It sometimes gives more information, as in the case of *bank* where the money-related senses of the noun are accompanied by three diamonds, whereas the water-related senses as well as the verb are given without a diamond. As this dictionary distinguishes subentries in a very limited number of clear homophones only[2], this is not a frequent practice. LDOCE4 systematically distinguishes between written and spoken data. In the case of *post*, for instance, the nominal senses belong to the 3,000 most frequent words in both written and spoken texts (marked as W3 and S3), but as a verb *post* still belongs to this list in spoken (S3), but not in written English (-). CALD2 characterizes part of the lexicon as **E** (essential), **I** (improver) or **A** (advanced). These marks are attached to entries or parts of entries which present a particular sense. For *watch*, for instance, the noun ('clock') and the verb ('look at') are marked as **E**, the phrasal verb *watch out* as **I** and the verb ('be careful') as **A**, whereas other senses and uses do not have any marking.

Thirdly, different numbers of elements are presented as belonging to the most important part of the lexicon. OALD7 and LDOCE4 both have a list of 3,000 words, Cobuild5 claims on its back cover to have "over 3,000 most frequent words in English clearly labelled" and presents (p. 1703-1704) a list of some 650 words having three diamonds. MEDAL2 has 7,500 "red words", i.e. words with one or more stars, of which 2,500 have three stars and are used as defining vocabulary. CALD2 presents 4,900 lexical elements as "essential", 3,300 as "improvers" and 3,700 as "advanced".

Table 2 presents some very frequent words as they are presented in the five learners' dictionaries. It gives the number of stars (MEDAL2) or diamonds (Cobuild5) that are given in these dictionaries. For OALD7, a "+" indicates that the word is included in the "Oxford 3,000 list of most important words". For CALD2, I have associated the "E" with the category of "3 stars", "I" with the category

[2] This distinction of homophones was not made from the start of the Cobuild dictionaries, where each form had only one entry even in cases where this form covered quite different words. This is still the case in many entries like *mass* where the "large number" senses are given alongside the "physical" sense and the "church ceremony" sense in one single lemma.

of "2 stars", and "A" with the category of "1 star". As for LDOCE4, in order to make the data comparable I have transformed the data in the following way: a "1" was changed into a "3" and a "3" into a "1"; in addition, I have taken the mean of written and spoken frequency data to calculate the number of "stars".

	OALD7	MEDAL2	Cobuild5	LDOCE4	CALD2
father n.	+	3	3	3	3
faucet	+	0	0	0	0
fault n.	+	3	2	1	3
fault v.	+	0	2	1	3
February n.	+	3	0	0	3
fence n.	+	2	1	1	2
ferry n.	–	1	0	0	2
festival n.	+	3	1	2	2
few det, adj.	+	3	3	3	3

Table 2. Frequency as indicated in the five learners' dictionaries

This table is an oversimplification of what can be found in the dictionaries because each dictionary not only has a different system of symbols but, as was already stated, also takes into account different numbers of frequent words. For instance, three diamonds are given in Cobuild5 to 650 elements only, whereas CALD2 has 4,900 elements marked as **E**(ssential). Moreover, as has been said already, the elements that are marked in Cobuild5 are not of the same nature as the ones that can be found in CALD2.

Nevertheless, Table 2 shows that there are clear cases where all dictionaries give about the same information: *father* and *few* are presented as belonging to the most important words of English in all cases. But there also are many cases where the frequency data are quite different. This is not only the case with *faucet*, which is included in the list of 3,000 most frequent or important words in one dictionary only, but also with a number of the other words in this table. The noun *fault* is given two diamonds in Cobuild2, which means that it belongs to the approximately 1,800 most frequent words of English, whereas in LDOCE4 it is qualified as S2 and W3, which means that it belongs to the lower end of the 3,000 words that are marked for frequency. As a verb, *fault* is again presented with two diamonds in Cobuild2 but has no star in MEDAL2, which means that it does not belong to the list of 6,700 frequent or important words. Something similar can be said of *fence* as a noun: in MEDAL2 it is somewhere in between 2,500 and 4,500, in Cobuild5 and LDOCE 4 in between 1,800 and 3,000, but in CALD2 in between 4,900 and 8,200. The noun *festival* is in a similar position. In some cases the heterogeneous indications about frequency are only seemingly contradictory, however. The noun *ferry* is presented as an **I** (improver; here noted as 2) in CALD2, with one star in MEDAL2, and with no marking in the other dictionaries. As these other dictionaries do not go beyond some 3,000 items, the absence of any marking is not inconsistent with the markings in the two other dictionaries.

The disparity of the frequency data shows again when two stretches of words, from *dad* to *decisive* and from *serious* to *shooting*, are compared. A total of 120 different word forms are marked in one or more of the five dictionaries as frequent or important. Only 48 of these words, that is 40%, are marked in some way in all five dictionaries, whereas 20 words (17%) are marked in only one dictionary and 19 words (16%) appear in two dictionaries. Most of the words that are marked in only one dictionary are found in MEDAL2 (14 words), the others are presented in CALD (3 words) and in Cobuild5 (3 words). MEDAL2 not only marks a number of unique words with one star, e.g. *dart***,** *decision-making*, *settler*, *shipment***,** but there are words that have two stars in this dictionary as well: *dealing*, *sexuality*, *shaft*. According to MEDAL2, these words belong to the intermediate category of most important words, but they do not have any particular marking in any of the other dictionaries. This is also the case with *seventeenth* and *seventieth*, which both have two diamonds in Cobuild5 and no particular marking in any of the other dictionaries. If we compare the marked words in the same stretches in OALD7 and LDOCE4, which both present a list of about 3,000 common words, we find that of the 82 words that have a marking in these two dictionaries, 24 (29%) of

them have contradictory markings, that is to say that they have a marking in one dictionary but not in the other. And again this does not only concern the least frequent ('1 star') cases; there are words like *database*, *debt* and *setting* that according to LDOCE4 belong to the intermediate ('two star') category but are not part of the OALD7 list. All these differences may be the result of choices made by the compilers, but it is hard to see how these choices can be motivated.

3. What is frequent in language?

Apart from the choice of other aspects that have played a role in the establishment of the lists of common words, these differences ask a more fundamental question: what is frequent in language? Although computers can most easily count forms, it is evident that these are not the most interesting elements in language use or in language learning. Language is used for meaning and a particular meaning can in most cases be rendered by different forms. At the same time, one particular form can in many cases correspond to a great variety of meanings. The form *favour*, for instance, which has a "key" in OALD7 because it is on the list of the 3,000 most important words, has in that dictionary five senses as a noun, four as a verb, and is part of four particular, idiomatic expressions. Not all these senses and uses are equally frequent or important; some of them are even quite infrequent, like the old-fashioned *sexual favours*. On the other hand, the mere fact that a given form has a great number of senses and uses is one of the reasons why it will be more frequent than forms that do not correspond to many meanings. As is well known, the most frequent words are almost always highly polysemous. In two samples taken in the four learners' dictionaries that give information about the relative frequency or importance of the marked words (that is: in all except OALD7), from *imply* to *indecision* and from *reproach* to *retire*, I counted a mean of 8.6 lexical units for the most frequent forms ('3 stars' according to the system explained above), 4.3 lexical units for the intermediate category, and 3.2 for the '1 star' category, whereas forms without any marking have a mean of 1.4 lexical units. Table 3 gives an overview of the figures I found for each dictionary. As can be seen, in MEDAL2 and LDOCE4, as well as in the mean, the numbers of lexical units covered by more frequent forms, those of the intermediate category, those of the '1 star' category, and those with no star, are in a strict descending order. For Cobuild5 and for CALD2 the figures for the '1 star' category are higher than those for the "2 star" category.

Forms marked with	MEDAL2	Cobuild5	LDOCE4	CALD2	Mean
3 "stars"	7.5	10.0	14.3	6.2	8.6
2 "stars"	3.6	6.6	5.0	2.8	4.3
1 "star"	2.7	8.0	2.8	3.3	3.2
no "star"	1.5	1.4	1.5	1.2	1.4

Table 3. Mean numbers of lexical units covered by lemma forms
(*imply* to *indecision*, *reproach* to *retire*)

What should be clear by now is that what has to be counted is not forms but lexical units, i.e. well determined linguistic forms that have a particular meaning and a well defined set of uses. For a learner it is important to know that *favour* may have the sense of "help" or "something kind", that in that sense it is used with verbs like *ask* or *do*, and that these particular uses are more frequent and, therefore, more important to learn than other senses or uses of the same form or of other forms. Corpus analysis has seen important developments over the years and is now able to yield this type of data (see, for instance, Fontenelle 2003 and Hanks 2004). A verb like *argue* has a number of different senses which appear in different constructions and use different prepositions (*to argue with someone, about something* or *to argue for something* (cf. Atkins et al. 2003). These different lexical units do not necessarily have the same kind of frequency. Moreover, it is not the cumulative frequency of these lexical units, which happen to share the same lemma form, that makes each of them more frequent or important than other lexical units. Even metaphoric uses of a word like *storm*, as in *a storm of protest* or *a storm of applause*, do not all have the same kind of frequency (Hanks 2004: 260) and may be more or less important for learners.

Not many of the results that modern corpus linguistics has made available concerning the marking of frequent senses or uses have yet made their appearance in the learners' dictionaries of English. In most of them, meanings are presented in order of descending frequency. But this does not always apply to

idiomatic uses, which may be included in separate sections of the lemma, after the list of meanings. In OALD7 and MEDAL2 the expression *do me a favour*, for instance, which is quite frequent in spoken language, is presented after the complete list of meanings of the noun, some of which are less frequent. LDOCE4 presents this expression as one of the ten lexical units, i.e. meanings or multi-word expressions, that are listed under *favour* as a noun, which seems to be a better solution. CALD2 sometimes marks an expression or a phrase as more important than other lexical units. Under *reputation*, for instance, which is marked as **E** (essential), we find *by reputation* marked as **A** (advanced), and under *feel* EXPERIENCE **E**, the expression *feel like sth* is marked as **I** (improver).

In order to complete the picture, it is useful to acknowledge the presence, in some of the dictionaries, of other types of information about frequency. MEDAL2 sometimes presents lists of words that are frequently used with particular senses of other words, like verbs that accompany the noun *stress* (e.g. *cause, create, manage, reduce*) or nouns that are frequently used as objects of the verb *stress* (e.g. *importance, need, urgency*). LDOCE4 provides graphs showing differences in frequency between American and British English (e.g. at *shop*, which is compared to *store*) or between expressions in written and oral texts (e.g. at *okay* or *a little* as against *a bit*), or providing frequency information on grammar patterns (e.g. at *stop*).

4. Conclusion

I have tried to make clear that in the most recent editions of the "big five" learners' dictionaries of English, more and more information about the frequency or importance of a core vocabulary is provided. This is done in quite diverse ways. A comparison between the data presented in different dictionaries casts some doubt on the reliability of these indications. Huge corpora as are used to compile the learners' dictionaries should yield more consistent information, not only about the one or two hundred most frequent forms or words of the language, but about several thousands of lexical units. As it is these lexical units, that is words, word senses or phrases, that learners of English have to acquire, these should be given due attention in the presentation of frequency information in this type of dictionaries. I think there is still room for improvement.

References

Dictionaries

[CALD2]. Walter, E. (ed.) (2005). *Cambridge Advanced Learner's Dictionary*. 2nd ed. Cambridge: Cambridge University Press.

[Cobuild5]. Sinclair, J. (ed.) (2006). *Collins COBUILD Advanced English Dictionary*. 5th ed. London: Collins.

[LDOCE4]. Summers, D. (ed.) (2003). *Longman Dictionary of Contemporary English*. 4th ed. Harlow: Pearson Education.

[MEDAL2]. Rundell, M. (ed.) (2007) *Macmillan English Dictionary for Advanced Learners*. 2nd ed. Oxford: Macmillan Education.

[OALD7]. Wehmeier, S. (ed.) (2005). *Oxford Advanced Learner's Dictionary of Current English*. 7th ed. Oxford: Oxford University Press.

Other literature

Atkins, S.; Fillmore, C. J.; Johnson, C. R. (2003). "Lexicographic Relevance: Selecting Information from Corpus Evidence". *International Journal of Lexicography* 16. 251-280.

Bogaards, P. (1996). "Dictionaries for Learners of English". *International Journal of Lexicography* 9. 277-320.

Bogaards, P. (2003). "MEDAL: a Fifth Dictionary for Learners of English". *International Journal of Lexicography* 16. 43-55.

Cruse, D. A. (1986). *Lexical Semantics*. Cambridge: Cambridge University Press.

Fontenelle, T. (2003). Special Issue of *International Journal of Lexicography* 16 (3) (on "FrameNet").

Hanks, P. (2004). "The Syntagmatics of Metaphor and Idiom". *International Journal of Lexicography* 17. 245-274.