Computers and the Humanities **34:** 99–102, 2000.

Lexicography and Disambiguation: The Size of the Problem

ROSAMUND MOON
Department of English, University of Birmingham, Great Britain

This contribution is by way of a footnote to discussions of the sense disambiguation problem, and it sets out quantifications based on the *Collins Cobuild English Dictionary* (hereafter **CCED**).[1] These suggest that as few as 10000–12000 items in the central vocabulary of English are polysemous: that is, potentially ambiguous at the word level. Other items have only one meaning in general or common use.

CCED contains 39851 different lemmas/headwords and headphrases.[2] It is intended for advanced learners of English as a second or foreign language, and its aim is to cover the central vocabulary of English as it appears in a large corpus of current text, the Bank of English (now 329 million words, but 200 million at the time of preparing CCED). While, clearly, no dictionary is either perfectly comprehensive or perfectly accurate, and while lexicographers produce at best only approximations of semantic and lexical truth, it has been assumed in the following that the analysis and coverage of meanings in CCED, because it is corpus-based, are a reasonable representation of those words and meanings found in general current English, and that the words and meanings not included in CCED are likely to be relatively rare, technical, or specialized in other ways. In fact, the headwords in CCED seem to account for about 94% of alphabetical tokens in the Bank of English, and over half the remaining tokens in the corpus are names, variant or aberrant spellings, and hapaxes.

Table I gives the number of headword items in CCED with particular numbers of senses. These headwords represent lemmas: CCED does not normally separate homographic forms into separate headword entries, even where they belong to different wordclasses or parts of speech, or where they are etymologically discrete. Accordingly, for example, nominal and verbal senses of *cut* or *dream* are treated together in single entries, as are *bark* 'noise made by dog' and *bark* 'outer covering of trees'.[3]

It can be seen that the majority of entries in CCED have just one sense, and only 5384, or 13%, have more than two. The average number of senses per dictionary entry is 1.73. If phrasal verbs, phrases, and idioms are excluded, the average number of senses per item is 1.84. Phrasal verbs each have an average 1.62 senses.

Table I. Headwords and senses in CCED.

Number of senses	Number of headwords	Proportion of headwords
1	27600	69.26%
2	6867	17.23%
3	2323	5.83%
4	1103	2.77%
5	591	1.48%
6–9	912	2.29%
10–14	289	0.73%
15–19	96	0.24%
20/over	70	0.18%

Phrases and idioms are generally not polysemous: 89% of those items included in CCED have only one sense, and the average is 1.06 senses.[4]

Many of the entries with two or more senses operate in two different wordclasses, and so are polyfunctional rather than, or as well as, polysemous: this means that the disambiguation process is simplified. If syntax as well as form is taken into account, a more refined assessment is possible. The next set of figures is based on a division of CCED entries according to wordclass, with nominal, verbal, adjectival, adverbial, phrasal, and other (miscellaneous) senses separated out. Closed-set items – determiners, prepositions, and conjunctions, such as *a, about, across, all, and* – have been excluded here for the sake of simplicity, and because the nature of the distinctions between their different 'senses' is generally syntactic, functional, or discoursal, rather than semantic: they are therefore not the primary targets for sense disambiguation work.

This division of CCED entries produces a new total of 49420 items, of which about 25% are polysemous. There are now in absolute terms more two- and three-sense items, since many of the heavily polysemous headwords in CCED have at least two senses in two or more wordclasses, but the average number of senses per item has fallen slightly to 1.69. Table II gives the profile for the whole set and for the specified wordclasses (the numbers of nouns etc. with two or more senses).

About 14% of the two-sense nouns can be disambiguated syntactically, through countability differences between senses, or formally, because one sense is capitalized: that is, pluralizability, determiner concord, and spelling are distinguishing characteristics. Although polysemous verb and adjective senses can sometimes be distinguished through transitivity, gradability, and prepositional or clausal complementation, this is comparatively infrequently a simple matter of binary distinctions: collocation and valency generally seem more significant criteria for lexicographers.

There is of course a close correlation between frequency and polysemy, and more heavily polysemous items are usually more frequent. The 455 headwords in

Table II. Headwords, senses, and wordclasses in CCED.

Number of senses	Number of items	Proportion of items	Nouns	Verbs	Adjectives	Adverbs	Phrasal verbs	Phrases
2	7362	14.9%	3513	1152	1616	369	357	242
3	2384	4.82%	1132	501	471	101	113	16
4	994	2.01%	444	263	153	44	57	2
5	527	1.07%	239	142	75	26	23	0
6–9	666	1.35%	247	196	111	20	34	2
10–14	163	0.33%	39	54	28	3	2	0
15–19	49	0.1%	5	17	11	0	0	0
20/over	33	0.07%	2	17	1	0	0	0

CCED with 10 or more senses all have at least 10 tokens per million words in the Bank of English, and together they account for nearly 50% of its alphabetical tokens. Many of the 455 are closed-set items: they are generally of very high frequency, alone accounting for 40% of the corpus. Many of the rest are versatile words occurring in many different collocations and contexts: CCED has used such features as criteria in making sense distinctions, even though there may be little substantial difference in core meaning. The most heavily polysemous of these items in CCED are the nouns *line, service, side, thing, time,* and *way*; the verbs *get, go, hold, make, run,* and *take*; and the adjectives, *dry, heavy, light, low, open,* and *strong*. At least some of these words are likely to have been finely split and undergeneralized in CCED in order to simplify explanations of meaning and to demonstrate typical lexicogrammatical and textual environments, for the benefit of CCED's target users. (See the paper by Krishnamurthy and Nicholls in this issue for a discussion of lexicographical procedures in relation to the Hector data.)

The above represents just one dictionary's account of polysemy and selection of headwords and meanings: other dictionaries of different sizes and types may provide different statistics. However, it may be used as a benchmark and as an indication of the extent of the task of sense disambiguation, whether manual or automatic. In this particular lexicon (approximately 40000 different entries, or 50000 if wordclass is taken into account), about 75% of items have either only one sense or only one sense per wordclass: nearly 1% more are closed-set items which do not need this kind of disambiguation. This leaves approximately 10000 of the headwords found in CCED (12000 items if separated into wordclasses) to focus on: about 7000 of these, in either case, have just two senses, and there are probably only 1000 very complex items to deal with.

Notes

[1] *Collins Cobuild English Dictionary* (1995, 2nd edition). London & Glasgow: HarperCollins.
[2] All data is copyright Cobuild and HarperCollins Publishers, and is reproduced with their permission.
[3] For convenience, exact counts are given here. These represent best attempts to extract the information from dictionary files, but variability and inconsistency in coding mean that other methods of counting could lead to slightly different figures. Note that the count of headwords/headphrases corresponds to 'lemmas', not to the dictionary publishers' conventional count of 'references', of which CCED contains 75000.
[4] Some of these items may be potentially ambiguous in another way, since identical strings with literal meanings can be found: for example, *in hot water* and *sit on the fence* can be used literally to denote physical location as well as idiomatically or metaphorically to denote non-physical situation or mental state. However, corpus studies suggest that this kind of ambiguity is relatively infrequent, and the institutionalization of an idiomatic meaning is typically associated with non-use of possible literal meanings.