Dictionaries for Language Learners

R. A. Carter and N. Schmitt

In the field of lexicography for langauge learners the most significant developments since the 1970s have involved more extensive corpora of spoken and written language and the creation of sophisticated computer-based access tools to such corpora. The greatest innovations have been stimulated by the COBUILD project at the University of Birmingham, England and the influence of such work can be measured by the fact that by the late 1990s all major English language learner dictionary projects have incorporated reference to extensive language corpora and developed computational techniques for extracting lexicographically significant information from such corpora.

1. The COBUILD Project

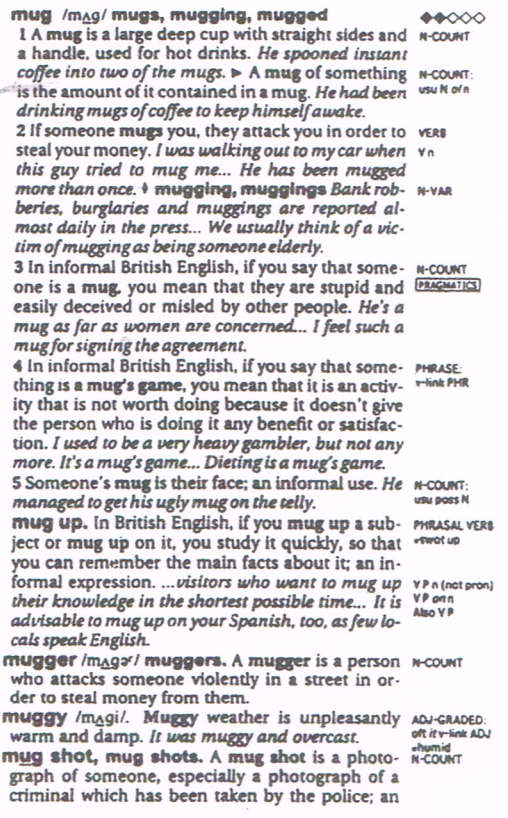
The COBUILD is one of the largest and most ambitious lexical research projects ever undertaken. COBUILD stands for Collins Birmingham University International Language Database and is based in the School of English at the University of Birmingham under the direction of Professor John Sinclair who, in addition to having major responsibility for lexical and lexico-grammatical research, is editor-in-chief of the major lexicographic and other related publications of COBUILD which began with the publication in 1987 of the ground-breaking *Collins COBUILD English Language Dictionary* (*CCELD*).

The principal aim underlying COBUILD research is to investigate in as much detail as possible how the English language is actually used at a given moment in time in both speech and writing and to allow such evidence to inform publications aimed at learners of the English language. As the project developed through the 1980s, it became clear that such evidence could only be made available by building a multimillion word corpus and the *CCELD* (1987) draws on a core database of 7.3 million words and makes supporting reference to a corpus of 20 million words. (For an account of early COBUILD corpus development see Sinclair 1987.)

Because most of the publications produced by COBUILD are for nonnative users of English, there has been less interest in the kinds of specialized one-off uses of language which are often of major interest in dictionaries for native users of English and correspondingly greater interest in the most central and typical uses of the language. Evidence in COBUILD dictionaries is therefore often given by illustrating meaning and usage in citations from the most typical and sometimes even the most banal examples of usage. The usage cited is corpus-based and includes real uses of English attested in actual, naturally-occurring usage and not therefore the made-up examples and citations of lexicographers which had characterized foreign language lexicography before 1987.

The main innovations of this first COBUILD dictionary and its latest edition the *Collins COBUILD English Dictionary* (*CCED*) (1995) can be summarized as follows and with reference to sample entries:

(a) Citations are examples of real English and do not involve made-up examples; the citations selected can be attested with reference to corpus evidence.
(b) Linguistic and stylistic differences between spoken and written usage and British and American English usage can be separately stored and marked accordingly in dictionary entries.
(c) Most crucially, relatively frequencies of occurrence are indicated and, most innovatively, in entries for individual lexical items the order of senses in multisense words corresponds to their frequency order in the corpus (Fig. 1).
(d) Concordancing techniques allow illustration of the main collocational and colligational properties of a word. Such properties can be made part of the explanation of a word's meaning. Significant lexical patterns and grammatical behavior are separately highlighted in *CCELD* in a separate column which is positioned in parallel with the relevant entry.
(e) Explanations are written in complete sentences



mug /mʌg/ **mugs, mugging, mugged** ◆◆◇◇◇
1 A **mug** is a large deep cup with straight sides and a handle, used for hot drinks. *He spooned instant coffee into two of the mugs.* ▸ A **mug** of something is the amount of it contained in a mug. *He had been drinking mugs of coffee to keep himself awake.* N-COUNT / N-COUNT: usu N of n
2 If someone **mugs** you, they attack you in order to steal your money. *I was walking out to my car when this guy tried to mug me... He has been mugged more than once.* ◆ **mugging, muggings** *Bank robberies, burglaries and muggings are reported almost daily in the press... We usually think of a victim of mugging as being someone elderly.* VERB: V n / N-VAR
3 In informal British English, if you say that someone is a **mug**, you mean that they are stupid and easily deceived or misled by other people. *He's a mug as far as women are concerned... I feel such a mug for signing the agreement.* N-COUNT PRAGMATICS
4 In informal British English, if you say that something is a **mug's game**, you mean that it is an activity that is not worth doing because it doesn't give the person who is doing it any benefit or satisfaction. *I used to be a very heavy gambler, but not any more. It's a mug's game... Dieting is a mug's game.* PHRASE: v-link PHR
5 Someone's **mug** is their face; an informal use. *He managed to get his ugly mug on the telly.* N-COUNT: usu poss N
mug up. In British English, if you **mug up** a subject or **mug up** on it, you study it quickly, so that you can remember the main facts about it; an informal expression. *...visitors who want to mug up their knowledge in the shortest possible time... It is advisable to mug up on your Spanish, too, as few locals speak English.* PHRASAL VERB: ▪ V P n (not pron) / V P on n / Also V P
mugger /mʌgə/ **muggers.** A **mugger** is a person who attacks someone violently in a street in order to steal money from them. N-COUNT
muggy /mʌgi/. **Muggy** weather is unpleasantly warm and damp. *It was muggy and overcast.* ADJ-GRADED: oft it v-link ADJ / ▪ humid
mug shot, mug shots. A **mug shot** is a photograph of someone, especially a photograph of a criminal which has been taken by the police; an N-COUNT

Figure 1. A dictionary entry for *mug*. *Source: CCELD*, 1987

(not in abbreviated phrases or codes) and involve a particular strategy of clear, accessible language (without recourse to a defining vocabulary) and a use of natural syntactic formulae. For example, 'if-clauses' are used for purposes of explanation, just as they frequently are in everyday discourse. Thus, lexical items are defined in context, often using the most frequent patterns which surround them in actual use, rather than as disembodied entities. A defining vocabulary is not employed but a note in the latest (1995) edition (*CCED*) states (p.xviii) that 'most words in our definitions [are] amongst the 2500 commonest words of English.'

(f) The COBUILD emphasis on the most frequent words in the language does not foreclose on the pragmatic or discourse functions of some of these frequent words. Thus, discourse markers such as *now*, *well*, *right*, and 'content-less,' propositionless words which have been largely ignored in previous dictionaries of this type, are also accorded illustration and explanation. A word such as *fine*, for example, is explained in a range of different senses but its meaning and function in conversational replies and as a marker of conversational boundaries as in: 'How are you? I'm *fine* thanks' or 'Is there anything anyone wants to add to this? OK, *fine*. Let's move on.'

In the years following the publication of *CCELD*, great efforts were invested in further corpus development as it was realized that lexico-grammatical description could be even better with a corpus with more words and more coverage from more different varieties of the English language. For example, Clear et al. (1996) note that in the 7.3 million word corpus there is only evidence that the word *taciturn* is used predicatively but the 20 million word corpus reveals that it is also used as a premodifier and regularly with another negative adjective as in *taciturn and unfriendly*. Descriptions were modified in the light of further evidence.

The COBUILD corpus, previously termed the *Birmingham Collection of English Test* (*BCOET*), was renamed *The Bank of English* in 1991 and in 1998 stands at 350 million words. The corpus has informed work on grammar and on idioms including the *Collins COBUILD Dictionary of Idioms* (1995), which gives unique guidance concerning both the frequency of different idioms and the different patterns which idioms form in varying degrees of fixedness and also including a dictionary of collocations *Collins COBUILD English Words in Use* (1997), which describes over 100,000 collocations in a range of lexical patterns and supported by attested examples from *The Bank of English*. Parallel publications include a series of concordance samplers for use in the classroom and CD-roms giving a wide variety of linguistic profiles of word usage. Simultaneously, the corpus is being continually updated to include a wider variety of spoken forms and data from other Englishes around the world.

The most substantial insight to have been generated by COBUILD research is that grammatical and lexical patterns are coselected and mutually interdependent. Clear et al. (1996) have expressed this as follows:

> Particular grammatical patterns tend to cooccur with particular lexical items, and—the other side of the coin—lexical items seem to occur in only a limited range of patterns. The interdependence of grammar and lexis is such that they are ultimately inseparable, working together in the making of meaning. (p. 313)

It is likely that future developments in lexicography will follow such insights; in the meantime it is worth reflecting that the publication of the first COBUILD dictionary in 1987 was greeted as idiosyncratic and unproven but that in the following 10 years corpus based lexicography following COBUILD lines has been adopted as standard practice in research, linguistic description, and publishing outcome.

2. Further Major Innovations

Other major and influential contributions to EFL lexicography have continued with subsequent editions of the *Longman Dictionary of Contemporary English* (*LDOCE*) (third ed. 1995) and the *Oxford Advanced Learner's Dictionary* (*OALD*) (1995). Although influenced by COBUILD computational methodology and, in particular, by the now established prerequisite of a corpus of linguistic evidence, subsequent innovations and developments have uniquely evolved according to different presentational principles.

In terms of corpora, both *LDOCE* and *OALD* have benefitted from the *British National Corpus* (*BNC*)—a corpus of 100 million words of written and 10 million words of spoken English—in the development of which both publishers (Longman and OUP) have been partners. Additionally, Longman has further extensive corpora of American English which inform all dictionaries including the *Longman Dictionary of American English*, the *Longman Lancaster Corpus* (*LLC*) (30 million words of written English) developed with advice from Professor Geoffrey Leech at Lancaster University and a 10 million word learner corpus including written texts from students at all levels from over 70 different language backgrounds and designed to provide evidence of the kinds of lexical mistakes most frequently made by learners as well as guidance concerning the kinds of words most likely to be understood by learners of English in dictionary definitions and explanations. Evidence from spoken corpora, in particular, has informed *LDOCE* (1995) in that the top 3000 most frequent words in speech (as opposed) to writing are marked out for special attention (see Fig. 2).

Other particularly characteristic features of *LDOCE* (1995) include:

(a) a continuing adherence to a finite defining vocabulary and to varied definition styles. (The 1995 edition of the Defining Vocabulary shows how the word list is being constantly revised in the light of research with users.) Another avowed aim is, where possible, to define the unit of meaning rather than individual words; this means that there are regular entries for phrases as well as for words. Selection restrictions on particular word forms are also clearly indicated;

(b) in *LODCE* (1995) a feature called 'signposts' is introduced to aid learners with the disambiguation of polysemous items. Signposts help the learner to make mental connections with the word in the context in which they encountered it;

(c) dictionaries and related materials are corpus-based but not corpus-bound. In other words, examples are given in an order which is most likely to help the learner rather than solely on the basis of the frequency of one sense rather than another. Authentic citations from the corpus are similarly not always helpful to the learner and in *LDOCE* it is an important principle that pedagogic mediation should precede the reality of the example.

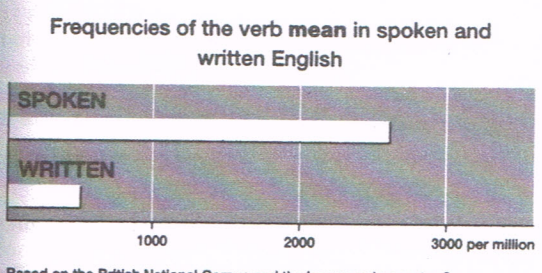


Figure 2. Frequencies of the verb *mean* in spoken and written English. *Source: LDOCE*, 1995

The fifth edition of *OALD* (1995) and the first edition of the *Cambridge International Dictionary of English* (*CIDE*) (1995) similarly contain numerous innovations. *CIDE* draws on the 100 million word Cambridge Language Survey (now the Cambridge International Corpus), with an emphasis on different national variations in English use and containing practical yet inventive features such as lists of false friends in English in comparison with 14 other international languages. *CIDE* also contains guide words which, in the case of polysemous words, orient the reader to the main or core meaning of the words listed in a single entry.

OALD (1995) represents a marked extension of a number of key features and some innovations in other areas, with the 1995 edition offering a treatment of 2800 new words and meanings when compared with earlier editions. Additional features include: 90,000 corpus based examples (drawn from the 100 million word *British National Corpus* (*BNC*) and the 40 million word *Oxford American English Corpus*); notes and illustrated pages giving information on cultural differences between British and American English; extensive usage notes covering areas of grammar and meaning which cause difficulty; and an expanded defining vocabulary (now 3500 words) is retained for purposes of definition and explanation.

3. Lexicography and English Language Learning: Contrasts and Comparisons

Table 1 summarizes some basic data about the four main learner's dictionaries, versions of which were all published in the year 1995. Comparisons between these dictionaries depend, however, on the criteria adopted for comparison and the grounds can never

Table 1. Some data about four learner's dictionaries of English

	LDOCE	*OALD*	*COBUILD (CCELD, CCED)*	*CIDE*
First edition (year)	1978	1948	1987	1995
editor(s)	P. Procter	A.S. Hornby	J. Sinclair P. Hanks	P. Procter
Latest edition/year	3/1995	5/1995	2/1995	1/1995
editor(s)	M. Rundell	J. Crowther	J. Sinclair G. Fox	P. Procter
No. of pages (a–z)	1644	1392	1951	1701
No. of other pages	64	78	38	91
No. of definitions claimed	>80 000	65 000	>75 000	1000 000
No of examples claimed	–	90 000	100 000	>100 000
Corpora	*LLC+BNC*	*BNC+OAEC*	*BE*	*CLS*

After: Bogaards 1966.
Abbreviations: LLC, Longman Lancaster Corpur (30 million words); *BNC*, British National Corpur (100 million words); *OAEC*, Oxford American English Corpur (40 million words); *BE*, Bank of English (200 million words; as of 1994); *CLS*, Cambridge Language Survey (100 million words).

therefore be entirely neutral, nor can any comparison be entirely valid without extensive empirical testing with users. However, among the evaluative frameworks to which reference needs to be made, at the very least according to the publishers' own criteria, are:

(a) clarity of definition and explanation and the extent to which defining vocabularies assist in this aim;
(b) authenticity, naturalness, and pedagogic mediation of examples;
(c) ease of access to the most frequent uses and core meanings (which are, of course, not necessarily identical),
(d) the extent to which words are shown in natural syntactic and collocational environments,
(e) the extent to which polysemous words and words which mean differently in different phrasal forms are appropriately explained and ease of access to them is provided.

A detailed comparison of these dictionaries is given in a special feature of *International Journal of Lexicography* (1996), (see in particular, Bogaards (1996) and Herbst (1996). See also Bejoint (1994) and Scholfield (1997)). Cameron (1998) raises valuable issues concerning the absence of diachronic information in many modern dictionaries, arguing that important cultural and ideological inflections become thereby deleted.

4. A Dictionary for Production

The *Longman Language Activator* (*LA*) (1994) is, uniquely, a production dictionary. It is aimed at intermediate to advanced learners of English and is designed around a conceptual map of the core words of English. These 1052 key concepts include words such as *sad/unhappy* around which are grouped, in a kind of atlas of meaning, a further 13 related words and phrases such as *be fed up with*, *be down in the dumps*, *depressed*, *miserable*, *downcast*, *glum*. These related words and their different levels of meaning and style are explained with reference to the core concept in such a way as to help students produce a range of expressions.

This information about meaning helps learners who know what they want to say but are seeking far more precise expressions; the learner should feel confident about expressing their ideas because information about a range or related meanings is given clearly and in accessible definitions (using a defining vocabulary). One aim of a production dictionary is to generate greater learner autonomy by encouraging learners to check, prior to use, how a word is used and in what collocational and colligational patterns. Decoding dictionaries involve, generally but not exclusively, less active modes of understanding. By contrast, the *LA* is essentially an encoding dictionary.

5. Bilingual Developments

Bilingual dictionaries have been extremely popular with language learners for a long time. In spite of this, second language teachers have viewed them with mixed emotions. On one hand, students could manage using them, being a quick and easy source of lexical information. On the other hand, it was clear that entries in bilingual dictionaries were often misleading and sometime simply wrong. What was needed was a

reference which retained the advantages of a bilingual dictionary, but which gave reliable information. Such a reference is now available with the introduction of the *Word Routes* (1995–). This innovative series combines easy lookup and accessibility with word entries compiled according to corpus-informed best practice. The explanations are in the first language (L1) but with numerous examples in the target language. Related words are clustered together in a thesaurus-like arrangement, with L2 and L1 indexes at the back to guide learners to the appropriate entry. Clustering in this way helps the learner to compare and contrast related words, and begin building the sense relationship connections necessary for native-like usage. There has never been any reason why bilingual dictionaries could not be compiled to the same standard and lexicographic good practice as the best monolingual ones; however, one disadvantage is that the cost of compiling a different dictionary for each different L1 means that only the major languages are ever likely to benefit from this development.

6. Conclusions

English language lexicography has undergone a phase of considerable invention and innovation in the last three decades of the twentieth century. A number of problems in the presentation of lexical information, particularly to language learners, have been solved and there have been considerable advances in the treatment of fixed and idiomatic expressions.

It is paradoxical that the most significant advances in the description of lexico-grammatical patterns have coincided with a time when the interests of linguists have shifted towards patterns of lexis in discourse. This means that lexicography is probably on the verge of even more exciting developments, including a major issue to address, in both theory and practice, in demarcating where grammars stop and where dictionaries start. For fuller surveys, reviews, and analysis see Carter (1998) and Schmitt and McCarthy (1997).

See also: Lexicology.

Bibliography

Bejoint J 1994 *Tradition and Innovation in Modern English Dictionaries*. Clarendon Press, Oxford, UK

Bogaards P 1996 Dictionaries for learners of English. *International Journal of Lexicography* **9,** 4:277–320

Carter R A 1998 *Vocabulary: Second Language Pedagogy: Applied Linguistic Perspectives*, 2nd ed. Routledge, London

Cameron D 1998 Dreaming the dictionary: Keywords and corpus linguistics. In: *Keywords: A Journal of Cultural Materialism, Vol. 1*, pp. 35–46

CCED 1995 *Collins COBUILD English Dictionary*. HarperCollins, Glasgow and London

CCELD 1987 *Collins COBUILD English Language Dictionary*. HarperCollins, Glasgow and London

CIDE 1985 *Cambridge International Dictionary of English*. Cambridge University Press, Cambridge, UK

Clear J, Fox G, Francis G, Krishnamurty R, Moon R 1996 COBUILD: The state of the art. *International Journal of Corpus Linguistics* **1,** 2:305–16

Herbst T 1996 On the way to the perfect learners' dictionary: A first comparison of OALD5, LDOCE3, COBUILD2, and CIDE. *International Journal of Lexicography* **9,** 4:321–57

LDOCE 1995 *Longman Dictionary of Contemporary English*. Longman, Harlow, UK

LA 1994 *Language Activator*. Longman, Harlow, UK

OALD 1995 *Oxford Advanced Learner's Dictionary*. Oxford University Press, Oxford, UK

Scholfield P 1997 Vocabulary reference works in foreign language learning. In: Schmitt N, McCarthy M (eds.) *Vocabulary: Second Language Pedagogy: Description, Acquisition and Pedagogy*. Cambridge University Press, Cambridge, UK

Schmitt N, McCarthy M J (eds.) (1997) *Vocabulary: Second Language Pedagogy: Description, Acquisition and Pedagogy. Cambridge University Press, Cambridge, UK*

Sinclair J (ed.) 1987 *Looking Up: An Account of the COBUILD Project in Lexical Computing*.

Word Routes (1995–) (ed. McCarthy M J) *English–Italian; English–French; English–Spanish*. Cambridge University Press, Cambridge

Lexicology

A. P. Cowie

This article is concerned with analyses of the vocabulary of a language (its 'lexis') in applied linguistics, and in particular with detailed descriptions intended to meet the needs of foreign learners. The principles discussed apply to all languages, though the examples are from English and the emphasis chiefly (though not exclusively) on the work of UK lexicologists. Though it has not ignored relevant theoretical developments, lexicology in the UK has taken a firm descriptive or empirical course. Its findings have appeared in various published forms: as dictionaries aimed directly at the foreign student (Cowie 1989a); as reports on lexical