Liz POTTER, Cobuild

Setting a good example. What kind of examples best serve the users of learners' dictionaries?

Abstract

Some have argued that learners are better served by examples that are, to a greater or lesser degree, invented by lexicographers than by examples selected from a corpus. This paper argues that learners are best served by carefully chosen corpus examples, not only because these represent the language as it is actually spoken and written, but also because learners can rely on the validity and accuracy of the information which the examples contain. It is argued that wholly or partially invented examples are not equally reliable reflections of usage, and that the large corpora available today give ample scope for finding suitable examples even for infrequent words and phrases.

Keywords: examples, learners' dictionaries

Since so many leading publishers are bringing out learners' dictionaries which claim to be corpus-based, it is clear that the analysis of corpora is an essential part of the modern dictionary-making process. However, those who write ELT reference works are not agreed on the use that should be made of the information corpora provide. In particular, there is disagreement over whether dictionary examples should be taken directly from corpora with little or no modification, or whether corpora should be used rather as the basis for invented examples; or indeed whether a combination of these techniques is desirable. (See for example the prefaces to the latest editions of LDOCE[1] and OALD[2], and to the CIDE[3]).

Over the past decade, the distinction between real and invented examples in English learners' dictionaries has become somewhat blurred. Before the first edition of the COBUILD dictionary (CCELD)[4] was published in 1987, practically all the examples that appeared in learners' dictionaries were made up, with the exception of occasional citations from magazines and newspapers. CCELD's innovation of using all real examples taken directly from a corpus led other dictionary producers to start using corpora as the basis, though not necessarily as the source, for their examples. So English learners' dictionaries now contain a range of example types from examples that are completely invented, through those that are partially invented but corpus-based, to those taken directly from a corpus with or without editorial modification.

Of course, even COBUILD dictionaries do not present totally unmediated chunks of corpus text as examples of usage. Many examples in the 1995 edition of CCED[5], as in all COBUILD dictionaries, were edited, especially for length (because of the inevitable restrictions of space imposed by the printed format) and to remove distracting, obscure or possibly offensive elements. However, it is still true to say that COBUILD dictionaries contain examples that have been taken directly from a corpus with a minimum of editorial intervention or alteration, while those in other learners' dictionaries range from completed invented to completely 'real', with many examples being corpus-based; that is, written after the lexicographer has consulted a corpus but not taken directly from it.

There is no doubt that using corpora as a regular part of the compilation process has brought about great improvements in the usefulness and naturalness[6] of the examples published in learners' dictionaries. More often than not, the examples now include one or more useful collocations and, where appropriate, a range of grammatical patterns. However, it is still not uncommon to find inaccurate, stilted, over-explanatory and unrealistic examples such as these: "The horse cocked (up) its ears on hearing the noise", "In the crash, the driver (was) catapulted through the windscreen" (OALD); "The artist cocked a snook at the critics by exhibiting an empty frame" (LDOCE); "To prove his skill as an acrobat he cartwheeled gracefully into the room" (CIDE). Examples such as the following, taken directly from The Bank of English, convey the flavour of the target words much more fully: "He suddenly cocked an ear and listened"; "So violent was the impact that the car was catapulted through the air"; "Miyako Yoshida proved the diversity of her talents, which extended to cartwheeling across the stage".

Some researchers (for example Laufer 1992[7], Nesi 1996[8]) have argued that examples invented by lexicographers are as useful or more useful to learners as those taken directly from a corpus with little or no modification. They argue this on the grounds that invented examples demonstrate the linguistic points the lexicographer wishes to convey, without any distraction or added difficulty such as may be introduced by using examples taken directly from real texts not produced for the purpose by the lexicographer.

Such findings contrast with the preliminary results of a recent survey of learners' dictionary users conducted by COBUILD, which found overwhelming approval among teachers and learners of English for real examples taken directly from a corpus. At the time of writing, 190 respondents had reported that they liked real examples or liked them a lot; while only 22 were indifferent or disliked them. A very few respondents made adverse comments relating to the fact that real examples are sometimes distracting or odd. Although this undoubtedly can and does happen, bigger corpora and more careful selection of examples should make it less of a problem. Moreover, the fact that these dangers exist is not an argument against taking examples directly from corpora or for inventing them either in part or in whole. Rather, it is an argument for using the resources offered by corpora in a careful and judicious way.

I would argue that corpora should be used not only as the basis for the lexicographer's analysis of the language but also as the direct source of the examples that are used to illustrate these findings. Having consulted a corpus in order to establish such matters as what collocates a word or phrase has, or the verb forms in which it is typically used, there seems to be little point in then inventing examples based on that information. The corpora available today are sufficiently large to enable lexicographers working on learners' dictionaries to find suitably clear and undistracting examples even for infrequent words and expressions. Furthermore, choosing examples from a corpus takes no longer than making them up, and may even be quicker, if the process of inventing the examples is preceded by corpus consultation and if the invented examples are to be sufficiently varied and interesting.

One situation in which a good case can be made for inventing examples, or at least radically simplifying those derived from corpora, is when they are designed for the use of pre-intermediate learners, who will find most unedited authentic text hard to cope with. However, the target users of the big EFL learners' dictionaries are at least at intermediate level and probably higher: they need that level of competence in order to understand the definitions, never mind the examples. Well chosen examples from up to date and varied corpora give

these users access to the very thing they are trying to master, the modern English language. It seems quite unnecessary to fob them off with a simplified version of it produced especially for that purpose by lexicographers, especially when the resulting examples can be unintentionally misleading, whether stylistically, grammatically, collocationally or in other ways (for example by simply not being 'natural' or plausible).

I would like to compare the results obtained by inventing examples (partially or entirely) as opposed to taking them directly from a corpus by looking first at an invented example produced by a fairly inexperienced lexicographer as part of an entry which I edited; and then at the examples for the same word in the four major English learners' dictionaries (CCED, CIDE, LDOCE, and OALD). The word being exemplified was the moderately common noun *bloom* (it occurs about 1300 times in the 323 million word Bank of English corpus). At first sight, *bloom* is a simple and straightforward count noun which means the same as '*flower*', but it is in fact stylistically very marked and is rarely if ever used in everyday speech and writing.

The invented example was: "I was just admiring the blooms in your garden". This is not a good example, but it needs a corpus to demonstrate exactly what is wrong with it. While a more experienced lexicographer would perhaps not have produced such a misleading example, even very experienced lexicographers cannot reliably produce accurate and helpful examples simply by a process of intuition and introspection. It is impossible to tell by introspection which adjectives collocate most frequently with a particular noun, for example, or that a verb is used predominantly in negative forms, or that the metaphorical use of a word is vastly more common than its literal use; all things that become evident immediately one consults a corpus.

"I was just admiring the blooms in your garden" is a well-formed English sentence: it contains a correct, though not very frequent, collocate of the target word (*garden*); it has the target word in the plural form, which in The Bank of English is more than four times as common as the singular; it shows a useful continuous verb structure and the typical subject-verb-object sentence order of English. And yet it is totally inadequate. No-one could argue that it is a better or more useful example than one taken from the corpus, because it is misleading in a way that a corpus example could not be. Furthermore, no-one using a corpus as the source of examples would ever come up with an example that misleads in the ways this one does, as it just would not occur.

One thing that is wrong with the example is that it suggests incorrectly that *bloom* is used in spoken English; in fact, it hardly ever is. Here are the frequency statistics for all the separate sub-corpora in The Bank of English, for about 1300 lines for *bloom/blooms*. (Forms which are obviously not nouns such as verbs, phrases etc have been excluded.)

Corpus	Total Number of Occurrences	Average Number per Million Words
brmags	567	18.8/million
brephem	49	10.4/million
oznews	140	4.2/million
brbooks	153	3.6/million
usephem	4	3.2/million
times	64	3.1/million
today	77	2.9/million
usbooks	90	2.8/million
usnews	22	2.6/million
guard	61	2.5/million
indy	43	2.2/million
newsci	8	1.3/million
econ	15	1.2/million
npr	11	0.5/million
bbc	8	0.4/million
brspok	2	0.1/million

The distribution among sub-corpora suggests that this use of *bloom* is stylistically very marked. It is used a great deal in specialised or semi-technical contexts (witness the high number of occurrences in sources such as gardening magazines in the British magazines sub-corpus). The occurrences in the ephemera sub-corpora tend to be in texts of a rather specialised nature such as catalogues for plant nurseries, or in the rather ornate and high-flown language of advertisements. Finally, there are a large number of occurrences in British and American books, which may be novels or non-fiction but are of course in all cases written sources. The clustering of **all** the spoken corpora at the bottom, at less than one occurrence per million words, suggests that this word is not at all common in spoken English.

As regards collocation, the evidence provided by The Bank of English suggests that *bloom* does indeed occur with *garden*, but not in the way shown in the made-up example. Typically it occurs in phrases such as "Their gardens are a mass of blooms and scents", or in much looser associations, such as this one from a British novel: "Spread across a scrub-top table in front of her were flowers from the garden, the last blooms of the late summer". Other items that *bloom* typically collocates with are colour words (*pink, white, yellow*); adjectives like *cut, single* and *double*; *faded, spent* and *dead*; or *delicate, beautiful* and *exotic*; nouns such as *plant, leaves, stem*, or *summer* and *winter*; and verbs such as *produce* and *fade. Bloom* does not typically collocate with *admire* (the combination occurs only once out of a total of about 1300 lines, the source being a letter to a gardening magazine). Not surprisingly, given that it was made up by the lexicographer on the sole basis of his intuitive knowledge of the English language, the invented example misses the opportunity to present the dictionary user with at least some of the useful information that is available about this word's behaviour, whether collocational, grammatical, or stylistic.

Suppose that instead of making the example for *bloom* up out of their head, the lexicographer were to do so after analysing the corpus data. They might conclude that the main features of the item are: that it is written, not spoken; genres in which it typically occurs are semi-technical or literary; the most frequent collocates include colour words, and so on. It would be impossible to invent a single example that showed all these features; ideally it needs at least

two, preferably accompanied by a list of significant collocates as well. To invent such examples the lexicographer would need to imitate, in this case, say the genres of gardening journalism and fiction. Even assuming that they were able to do this, why should they, when a large corpus immediately yields examples such as: "Carnations will produce a fine display of superb blooms in the late summer and early autumn"; "The medium-sized blooms are pink with clear stripes of crimson and purple"; "...a most attractive rose with bright green leaves and most beautiful large colourful blooms"; "After flowering, cut off the dead blooms above a healthy pair of buds" (all from the British magazines corpus); "..a beautiful creeper heavy with fragrant blooms" (from the Australian newspapers corpus); "He held up a single, perfect bloom. 'For you, to remember me by when we're apart'" (from the British books corpus).

Moving on to the examples for this sense of *bloom* in the published dictionaries, many inadequacies and omissions become apparent. In several cases, it seems unlikely that a corpus was consulted when writing the entry, or at any rate that corpus evidence played much part in the production of the examples. In the first place, only CCED gives accurate style and register information, observing that this word is "a literary use or a technical use in gardening". Of the other dictionaries, the CIDE marks one of its examples as literary, while the others give no indication that this word is restricted. (The same is true of the five bilingual dictionaries I consulted, all of which gave a straight translation equivalent to '*flower*', and gave no style or register warnings, and no examples). There is nothing to stop a learner who consults one of these dictionaries from coming out with a completely inappropriate utterance such as "What lovely blooms" or, indeed, "I was just admiring the blooms in your garden".

Here are the examples from the four learners' dictionaries, in alphabetical order:
"...a mass of bloom on the apple trees": the flowers you get on fruit trees are not normally called '*bloom*', but '*blossom*'. The phrase "mass of bloom" is very uncommon in The Bank of English, occurring only twice with both occurrences coming from gardening magazines. "...an exotic bloom": this example gives one good collocate, but with too little context to be really useful. "...beautiful red blooms": two good collocates, but again too little context. "Harry carefully plucked the bloom": unfortunately this example from The Bank of English was cut for reasons of space and in the process has lost much of its typicality. "The house had been filled with sweet-smelling blooms": this example is correctly labelled as literary, but the opportunity to present common collocates has been missed. Both *scented* and *fragrant* collocate more commonly with *bloom* than *sweet-smelling*, a collocation for which there are no lines in The Bank of English. "Their garden was full of wonderful blooms": this is very similar to the invented example I rejected. It is over-informative and stilted, and stylistically inconsistent - the natural choice for the last word would be something like '*flowers*' or '*plants*'.

"These chrysanthemums have beautiful blooms": this example seems very forced and unnatural. In The Bank of English, *beautiful* collocates with *bloom*, but not in conjunction with *have*: there are no lines for plants having beautiful blooms. *Bloom* tends to be used to refer to rather more fragile and short-lived flowers than chrysanthemums; and, as in the previous example, too much additional information has been included. A learner who used this as a model would risk producing language that was grammatically correct but not natural or idiomatic. "The sweet fragrance of the white blooms makes this climber a favourite": this example - from The Bank of English - comes from a very typical context for *bloom*, a gardening magazine. *White* is the commonest colour collocate; *fragrance* is a less common

collocate than *fragrant*, but it is in the right general area; *sweet* also collocates with *bloom*, although not all that frequently.

The aim of all dictionary producers is to provide their users with reliable and useful information about the language they are studying. Much of the information given in made-up or corpus-aided examples is not reliable either about the contexts in which a word or phrase is typically used, nor about the words that typically occur with it; and in some cases it can be actually misleading. At the very least, examples chosen from a corpus give the learner the guarantee that a piece of language actually does occur. If the examples are well chosen, they can and do give the learner a lot more than that.

Notes

1 *Longman Dictionary of Contemporary English*, 3rd Edition (1995) Addison Wesley Longman. Harlow.

2 *Oxford Advanced Learner's Dictionary*, 5th Edition, (1995) Oxford University Press. Oxford.

3 *Cambridge International Dictionary of English*, (1995) Cambridge University Press. Cambridge

4 *Collins COBUILD English Language Dictionary*, 1st edition (1987). Collins. London.

5 *Collins COBUILD English Dictionary*, 2nd Edition (1995). HarperCollins. London.

6 The concept of 'naturalness' is a complex and problematic one that has been considered by Sinclair and others. (Sinclair, J.M. (1988) Naturalness in language, in *ELR Journal* Vol 2, Birmingham, UK). In using the word here, I mean simply whether a particular piece of language is likely to occur or not, whether it would strike a native speaker as 'natural'. Sentences such as the examples for *cartwheel* and *cock a snook* are unnatural in that they are extremely unlikely ever to occur, except as illustrations for those particular words or phrases.

7 Laufer, B. (1992). Corpus-based versus lexicographer examples in comprehension and production of new words. Euralex '92. *Proceedings I-II. Papers submitted to the 5th Euralex International Congress on Lexicography in Tampere.* Finland. Part I, Studia Translatologica ser. A Vol. 2, 71-76.

8 Nesi, H. (1996). The role of illustrative examples in productive dictionary use. *Dictionaries*, 17; 198-206.