Proceedings of the XXII International Conference of AEDEAN, 111-116. Eds. P. Gallardo i E. Llurda. Lleida: Universitat de Lleida, 2000

Lexical Description in a Corpus-Based Dictionary and Grammar

Janet DeCesaris
Josep Maria Fontana
UNIVERSITAT POMPEU FABRA

This article is a critique of recent attempts to use corpora and the methods associated with corpus linguistics in creating pedagogical materials on lexical structure. The authors argue that the main reason for the inadequacy of these materials is the lack of realization that a corpus by itself does not provide all the information required to make claims about productivity. Using a corpus can be advantageous in lexicography because the intuitions of a single lexicographer, or even a group of lexicographers, may not accurately reflect current usage. Thus, consulting a well-structured corpus certainly ensures that any study previous to the elaboration of pedagogical materials will cover the right terri-tory. Such consultation, however, does not automatically result in linguistically significant generalizations, which are ultimately the key to useful grammatical descriptions, for the native and nonnative alike.

1. Introduction

The use of large corpora has had an enormous impact on lexical studies ranging from computational lexicons for machine translation projects to determining the productivity of a specific affix (Baayen & Renouf 1996, 82-94). Its impact on pedagogical lexicography has been particularly great in English because of the tremendous commercial success of the *Collins Cobuild English Dictionary* (hereinafter *Cobuild* dictionary), which in little more than ten years is already into its second edition. The information provided by the *Cobuild* dictionary is welcomed by the English language teacher and learner alike because the entries go well beyond noncontextualized efinitions to provide readers with usage information. What we might call "the *Cobuild* approach" was then extended to produce reference grammars and guides to specific areas of English, in order to make more detailed information available on a smaller scale. This series of materials, however, does not seem to have gained as wide an audience as has the dictionary. This, of course, might be due to extralinguistic factors such as marketing strategies, competing materials on the market, specific recommendations made by educational authorities favoring one type of pedagogical approach over another, etc., all of which are beyond the scope of academic analysis. It is, however, legitimate to subject the use of corpora in creating pedagogical materials on lexical structure to academic analysis. In this paper we suggest that the nature of the lexical information in a dictionary is rather different from the kind of information that is useful to a non-native speaker in a grammar guide, and that one important consequence of this difference is that a corpus-based description with little-to-no accompanying linguistic analysis, while passable as an approach towards lexicography, is simply not sufficient for a grammar. Using a corpus can be a useful tool in

organizing lexical material, but it is not a substitute for linguistic insight.

2. Corpus-based examples in pedagogical dictionaries

In this paper we will not go into a detailed review of what the *Cobuild* dictionary does and does not do; such information is available in Bogaards (1996), where the dictionary is compared with other dictionaries for non-native English speakers. What is important for the present paper is to point out that the defining style and examples that are characteristic of the *Cobuild* dictionary generally provide the reader with usage information. Complete sentences, often written in the second person, are used in definitions, as opposed to the brief, synonymic style of more traditional monolingual dictionaries (addressed to native speakers). The stated goal of *Cobuild*'s defining style is to set out the meaning "in the way one ordinary person might explain it to another" (John Sinclair, preface to *Cobuild* II, page xi). The result of this approach can be seen in the following definitions, as compared to a more traditional style of definition, in this case from the *American Heritage Dictionary* (*AHD*):[1]

(1)

Cobuild Dictionary (1995):
preconceived. If you have **preconceived** ideas about something, you have already formed an opinion about it before you have enough information or experience. *Five minutes after he had arrived for the interview, I had abandoned my preconceived ideas about boxers... We all start with preconceived notions of what we want from life.*
inappropriate
1. Something that is inappropriate is not useful or suitable for a particular situation or purpose. *The industry is inappropriate to the region's present and future needs... ...*
2. If you say that someone's speech or behaviour in a particular situation is inappropriate, you are criticising it because you think it is not suitable for that situation. *I feel the remark was inappropriate for such a serious issue... It is inappropriate for a judge to belong to a discriminatory club.*

American Heritage Dictionary (1992):
preconceive *tr. v.* **-ceived, -ceiv·ing, -ceives.** [the reader must deduce that the past tense/participle form ending in *-ed* can be used as an adjective] To form an opinion or a conception of (something) before possessing full or adequate knowledge or experience.
inappropriate Unsuitable or improper; not appropriate.

Although *Cobuild* does not explicitly state that *preconceived* often occurs in conjunction with the nouns *ideas* or *notions* as opposed to other nouns from the same semantic field, the examples effectively give the reader that information. Likewise, the definition for *inappropriate* is wisely separated into two senses, so that readers realize that the word can be used both in a more neutral, objective statement (sense 1) as well as in a clearly disapproving sense (sense 2).

The reason why definitions like these, which do not appeal to any specific labels such as **usage** or **pragmatics**, are useful for non-native speakers is that the contextualized examples illustrate usage. This use of corpus-based information entails being able to choose the best examples from those present in the corpus; the catch, of course, is determining what "best examples" in lexicography means. In pedagogical dictionaries issues such as subcategorization, collocations, frequency of a particular syntactic structure, and presence in an idiom are factors to bear in mind when choosing examples (Bogaards 1996, 280-1).

1. Some examples have been deleted from the entries to save space.

3. Corpus-based information in a word-formation handbook

The *Cobuild* handbook on word-formation includes information on approximately 300 prefixes and suffixes. The affixes are listed alphabetically, and the handbook attempts to give readers information on productivity by including a heading "productive use" under some entries (for example, under *ever-* or *-first*). A list of words formed with the affix as described in the definition is given under each entry to exemplify use. In contrast with the contextualized examples in the dictionary, the bulk of examples in this handbook are simply lists of words. The second sense for the prefix *pre-* (pp. 141-2) illustrates the type of entry in this guide:

pre- 2 Already

PRODUCTIVE USE: **pre-** combines with nouns and past participles to form new nouns and adjectives. Words formed in this way refer to or describe an action which has already been done. For example, a "preconception" is a belief that you already have about something before you know enough about it to form a fair opinion of it; if something is "prepaid", it has already been paid for. [Information about spelling is provided] [6 illustrative sentences]

Here are some examples of words with this meaning: [24 words listed]

pre-arranged predestination premeditation pre-planned (etc.)

Words with other meanings: [16 words listed]

preamble predominate prehistoric prepossessing (etc.)

This type of presentation of morphological information raises two important issues in the teaching of word-formation strategies: lexicalization of derived words and representation of productivity. The *Cobuild* corpus-based approach to derived words seems to be based on the assumption that a derived word is primarily the sum of its parts, and therefore a list grouping together words with the same formative and relatively similar meanings for that formative will be helpful to the language learner. Any word with the formative in question that does not show the predicted meaning is relegated to the "Words with other meanings" list. This approach, however, represents a simplistic view of word meaning that is not borne out by careful consideration of data. Derived words often acquire lexicalized senses, to the extent that one or more senses of a morphologically complex word can no longer be considered strictly compositional. In fact, what could be called "meaning drift" is rather commonplace with derived words, is well attested in dictionaries, and may affect both monosemous and polysemous words. Nominalizations with the suffix *-tion* are a good example of this phenomenon. For example, according to the Cobuild dictionary the verb *situate* is a synonym for *locate*, whereas the noun *situation* is commonly used to refer to a state of affairs that is happening, which would account for the fact that situation is not the usual deverbal noun for *situate:*

(2)

a. I was unable to situate them because so many years had passed.

b. *My situation of them (Acceptable: My situating them) was made difficult because so many years had passed.

According to the corpus, the most frequent meaning of *situation* is not 'location,' which nonetheless is the basic meaning of the verb; yet, *situation* is listed among the words in the word-formation guide as an example of a deverbal noun ending in *-ion*.[2] We note that this phenomenon of lexicalization can also occur with a particular sense of a given word; thus, *examination* can be interpreted as a deverbal noun derived from *examine*, as shown in (3):

2. The guide lists the forms under *-ion* in order to group together words ending in *-ation*, *-sion*, *-tion*, and *–ition*.

(3)

a. The committee examined the candidates' résumés

b. The committee's examination of the candidates' résumés

The word *examination*, however, is also used in the specific sense of a formal test to display knowledge, and this sense is also related to the idea *examine*, although the use of the verb *examine* in this context is questionable at best:

(4)

a. The college entrance examination will be held next Saturday.

b. * I was examined last Saturday.

Acceptable: I took the college entrance examination(s) last Saturday.

This lexicalization of a sense of *examination* does not mean that the word is morphologically irregular, but rather that the meaning is not compositional because it has acquired reference. Including words like *situation* and *examination* in a list of deverbal nouns with no additional comments is misleading because their interpretation is not always as clearly deverbal as the listing would sugggest.

Entries in the Cobuild dictionary show that non-compositional meaning can also be a characteristic of monosemous derived words. For example, the meaning of *pre-packaged* "**Pre-packaged** foods have been prepared in advance and put in plastic or cardboard packages before they are sold. *...pre-packaged duck and orange sauce.*" This meaning cannot be derived simply from either of the meanings of *pre-* plus that given for *packaged* (sense 4 under *package* "When a product is packaged, it is put into packets to be sold. *The beans are then ground and packaged for sale as ground coffee.*). Should this necessarily imply that the *pre-* in *pre-packaged* is unrelated to the prefix *pre-* as described? We believe not, but that appears to be the conclusion reached on the basis of the corpus material. The intricacies of the meaning of morphologically complex words can only be discussed in the light of linguistic analysis; lists based on forms with no reference to how the forms are used in context does not appear to provide valuable information.

The issue of productivity is no less problematic. Productivity of an affix cannot be determined by corpus data alone; as Baayen and Renouf (1996, 73-78) point out, the data from a corpus must be compared with that from a different point in time (either from a corpus or a dictionary) in order to make claims about productivity. Exactly what constitutes productivity, however, is not straightforward. Subregularities in morphology, such as the alternations exhibited by *receive-reception*, *perceive-perception*, *deceive-deception* are usually classified as productive although they may apply to a closed set of lexical entries (Carstairs-McCarthy 1992, 50). This would not seem to be the idea of productivity in the *Cobuild* series, as they editors explicitly refer to combinations "with a large number of words" (viii) and to the creation of new lexical items. The handbook, then, effectively divides affixes into two groups: those which display productive use, and those which do not.

It is uncontroversial that some affixes in English are productive whereas others are not (Bauer 1983, 62-100); the usefulness of the specific description provided in the *Cobuild* book, however, is not as clear. Space limitations limit our comments to a representative sample of affixes (those beginning with *m-*). The affixes in (5) are listed as having a "productive use", while those in (6) are not so described:

(5)

prefixes: mega-, mid-, mini-, mock-, much-
suffixes: -made, -minded

(6)

prefixes: macro-, matri-, micro-, mis-, mono-
suffixes: -mate, -meter, -most

The above division takes no account of neoclassical formatives, although the standard description of English word-formation (Bauer 1983) makes such

a division; a view of word-formation incorporating this class of elements would likely group *mega-* together with *macro-* and *micro-*, for instance. Identifying *matr-/matri-* as a prefix in English and putting it on a par in terms of productivity (or, as presented in this book, in terms of the lack thereof) with a prefix like *macro-* is misleading in that *macro-* is surely productive in scientific and technical language, whereas *matr-/matri-* is not productive at all. Furthermore, even if the neoclassical elements are removed from (5) and (6), several of the prefixes/suffixes listed above would not be considered affixes at all in an analysis of English word-formation (e.g. *much-*, *-made*, *-minded*, *-most*) but rather would be treated as words typically used in compounding. In short, the claims about productivity based solely on a corpus classify elements together that do not share similar linguistic behavior, and we wonder how nonnative speakers would be able to "make new words of their own" based on such groupings.

4. Conclusions

While the use of the corpus data in the dictionary yielded positive results, the attempt to use the same data as a basis for a description on word-formation proved much less successful. We would like to suggest why that should be the case. First, a corpus by itself does not provide all the information required to make claims about productivity because it is simply a snapshot of the language. A corpus, no matter how large, fundamentally gives you distributional data. But then again, that alone is not enough to explain word-formation processes in English. Using a corpus can be advantageous in lexicography because the intuitions of a single lexicographer, or even a group of lexicographers, may not accurately reflect current usage. Although good morphological analysis tends to make for better dictionary entries, in lexicography it is perhaps more important to make available a large amount of information to the reader. Thus, consulting a well-structured corpus ensures your dictionary covers the right territory, because the usage is set out for you. But such consultation does not automatically result in linguistically significant generalizations, which are ultimately the key to useful grammatical descriptions, for the native and nonnative alike.

Works cited

American Heritage Dictionary of the English Language, 3rd edition. Boston: Houghton Mifflin, 1992.

Baayen, R. Harald. & Antoinette Renouf. "Chronicling the Times: Productive lexical innovations in an English newspaper." *Language* 72 (1996): 69-96.

Bauer, Laurie. *English word-formation*. Cambridge: Cambridge UP, 1983.

Bogaards, Paul. "Dictionaries for Learners of English." *International Journal of Lexicography* 9.4 (1996): 277-320.

Carstairs-McCarthy, Andrew. *Current Morphology*. London: Routledge, 1992.

Collins Cobuild English Guides 2: Word Formation. London: HarperCollins, 1991.

Collins Cobuild English Dictionary, 2nd edition. London: HarperCollins, 1995.