CHAPTER

13

The Impact of Corpora on Dictionaries

Patrick Hanks

This chapter discusses how corpus-linguistic techniques have revolutionized dictionary creation since the 1980s. While arguing that corpora enable improved dictionaries, I address a number of issues which suggest that corpora should not be used unthinkingly, for example it is important for compilers to address questions such as whether a dictionary is intended primarily for decoding or encoding purposes, hence a corpus ought not to be used just to produce larger and larger new editions of dictionaries with more and more 'authentic' examples. Instead, corpus techniques should help dictionary creators to consider which words (or uses of words) should be left out of a dictionary (particularly if the dictionary is aimed at learners), and examples should be carefully and sparingly selected to illustrate normal usage. Additionally, I discuss the contribution of corpus approaches to lexicographic treatment of pragmatics, phraseology and grammar. The chapter ends with a brief look at research on the Pattern Dictionary, which is being compiled with evidence from the British National Corpus.

13.1 Early Corpora

Early electronic corpora, in particular, the Brown Corpus (Francis and Kučera 1964) and the LOB Corpus (Johansson et al. 1978) had little impact on lexicography, despite being consulted by some major dictionaries during the earliest days of corpus linguistics (in particular the *American Heritage Dictionary*, first edition, 1969; and the *Longman Dictionary of Contemporary English* (LDOCE), 1978). With the benefit of hindsight, the reason for this lack of impact was simple: these pioneering early corpora were not large enough to show significant facts about the behaviour of most individual words. They only contained one million words, so it was difficult to distinguish statistically significant co-occurrences of words from chance co-occurrences. The set of word forms in a language is not a fixed number, but we can estimate that something in the order of 250,000 types (unique words) are in regular use in English at any one time. Even allowing for Zipf's law (Zipf 1935) in relation to the distribution of words in a corpus – a phenomenon which can be crudely characterized as: 'most words occur very rarely; a few words occur very often', a corpus of only 1 million words has no chance of showing the user statistically significant collocations of any but a few very common individual items. In such a corpus, a few significant collocates for function words such as

prepositions can be detected, but some perfectly ordinary words do not occur at all, and for those that do occur, their collocations with other words cannot be measured effectively. In small corpora, almost all of the co-occurrences appear to be random even if they are not. Similarly, for most mid-to-low frequency words, a corpus size of only a million words does not give reliable information about the extent to which a word has multiple meanings or belongs to multiple grammatical categories.

It was left to a few pioneers in corpus linguistics, notably Francis and Kučera, Sinclair, Leech, and Johansson and Hofland, to struggle on undaunted for almost 30 years in the face of misguided and sometimes virulent hostility from the dominant 'generative' school of linguistics, whose adherents arrogated to themselves the term 'mainstream' (though 'backwater' might now seem a more appropriate metaphor). The research method of these generative linguists characteristically relied almost entirely on the invention of data by introspection, followed by some explanation of whatever it was that had been invented. Though always suspect (being in danger of trampling unwittingly over some constraint of naturalness or idiomaticity), the invention of data may be regarded as unexceptionable when used to illustrate simple, normal structures of a language. However, the programme of generative linguistics was in many cases to discover a sharp dividing line between syntactically well-formed and syntactically ill-formed sentences. One of the important discoveries of corpus linguistics and corpus-driven lexicography has been that no such sharp dividing line exists. There is an infinitely large body of obviously well-formed sentences and an infinitely large body of ill-formed sentences in a language, but there is no sharp dividing line between them. Skilled language users often deliberately exploit the conventions of normal usage for rhetorical and other effects. For this reason, when a dictionary user (in particular, a foreign learner) asks, 'Can you say X in English?' the lexicographer is constrained to provide answers in terms that assume that the question really is, 'Is it normal to say X in English?' The boundary between possible and non-possible use of each word is always fuzzy; conventions are always open to exploitation.

In a prescient paper, published as early as 1966, John Sinclair argued that an essential task for understanding meaning in language would be the analysis of collocational relationships among words, which 'would yield to nothing less than a very large computer'.

13.2 Corpus-Driven Lexicography: From Cobuild to MEDAL

Things began to change with the first edition of Cobuild (1987). This was specifically designed as a tool to help foreign learners of English to write and speak natural, idiomatic English. In other words, it was designed as an encoding aid rather than a decoding aid. In 1983, after long struggles, both with issues such as rights and permissions and technical issues such as how to handle such a large corpus on the University of Birmingham's computer, a corpus of 7.3 million words was completed (tiny in today's terms, but more than seven times the size of any previous corpus), This was used as a basis for compiling the first draft of the dictionary.