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.

The corpus yielded many new insights, large and small, but, with a corpus of only 7.3 million words, the lexicographers still allowed themselves to supplement its evidence by a combination of the evidence of their own intuitions and other dictionaries. By the time the final editing stage was reached (1986), the Birmingham Corpus had grown to 18 million words and lexicographers became rather more reluctant to trust their own intuitions in defiance of the absence of corpus evidence. In today's world, with corpora of hundreds of millions and even billions of words being available (see Culpeper, this volume), it is a foolhardy linguist or lexicographer who prefers his or her intuitions over very large samples of evidence, but that does not stop some people. Some generative linguists, cognitive linguists and construction grammarians (whatever their other merits) continue to blithely invent evidence purporting to demonstrate idiomatic uses of language where little or no empirical evidence exists. I recently received an email from an able and respected linguist (a non-native speaker) requesting advice on access to a corpus that would provide examples of constructions such as *I walked the letter to the post*, an example invented by Langacker which, not surprisingly, she had failed to corroborate in the British National Corpus (BNC). The possibility had apparently not occurred to her that Langacker's invented example is not idiomatic, and that her failure to find it in BNC might itself be an interesting piece of empirical data. Of course, lexicographers must be aware of the dangers of the failure-to-find fallacy (the fact that something is not found does not mean that it does not or cannot exist), but failure to find a phrase in a very large corpus suggests at the very least that it is not very idiomatic.

As an example of the new insights into word behaviour yielded by early work in corpus lexicography at Cobuild in the 1980s, consider the case of *–ly* adverbs. It was widely assumed by pre-corpus lexicographers that all (or almost all) *–ly* adverbs were adverbs of manner modifying the sense of a verb, an adjective, or another adverb and that the meaning of the adverb was always (or almost always) systematically derivable from the root adjective. There is some truth in this, of course: *walking slowly* is walking in a slow manner. But some *–ly* adverbs in English have special functions or constraints, which were not always well reported in pre-corpus dictionaries. For example, The *Oxford Advanced Learner's Dictionary* (OALD1-4), says nothing about the use of words like *broadly*, *sadly*, *unfortunately*, *luckily* and *hopefully* as sentence adverbs – linguistic devices that enable speakers and writers to express an opinion about the semantic content of what they are saying. Those pre-corpus dictionaries which did notice sentence adverbs did not succeed in noticing all of them systematically. They tended to be more concerned with questions about prescriptive rules, for example whether it is correct to say '*Hopefully, he will deliver his paper before the deadline.*' The problem here is that although '*sadly, he died*' can be paraphrased as '*It is sad that he died*', it is not the case that '*Hopefully, he will deliver*' can be paraphrased as '**It is hopeful that he will deliver*'. Corpus lexicographers now recognize that such concerns are based on a theory of language that assigns too great a role to lexical compositionality and too small a role to the idiosyncratic conventions that are associated with each word. The lexicon is indeed, in Bloomfield's phrase, a 'basic list of irregularities'.

However, whereas for Bloomfield (1933) and for Chomsky (1981) this was a reason for shying away from any attempt to show how words and meanings are related, for corpus lexicographers this is the central issue to be investigated. In the course of investigating such irregularities, new regularities are discovered. By the time of Crowther's corpus-based 5th edition, the OALD included a note at *hopefully* explaining adverbs which 'can modify the whole sentence'. For most (but not all) of them, an example (but not a definition) is given illustrating such use, for example at *sadly*: *Sadly, we have no more money.*

A glance at corpus evidence, even in quite a small corpus, shows the function of sentence adverbs, enabling a speaker to express their own attitude to the propositional content of what they are saying. It is a truism that '*Sadly, he died before completing the project*' does not mean that he died in a sad manner, but this was not always apparent to dictionary makers until they were confronted with overwhelming evidence.

Another example of what the early corpus evidence showed the Cobuild lexicographers is the fact that the supposed adverbs of manner – even those that really are adverbs of manner – do not always and regularly inherit the entire semantics of the root adjective. The adjective *lame*, for example, has two senses: 'unable to walk or run properly because of an injury to a leg' (applied to animate beings) and 'disappointingly feeble' (applied to excuses and other speech acts). The corresponding adverb, *lamely*, on the other hand, very rarely has the 'injured leg' sense. It is almost always used in the 'feeble excuse' sense. In an informal experiment, students at Birmingham University were asked to invent a short anecdote, ending with the sentence, 'She walked lamely out of the room'. The majority of them invented a story in which the person concerned felt that she had no adequate reply, rather than one in which she had injured her leg. This informal experiment needs to be repeated under better controlled experimental conditions, but it suggests that that even collocation with the verb *walk* is not always strong enough to activate the 'injured leg' sense of *lamely*. This is not a necessary condition for the idiomatic use of lamely – the BNC contains two or three examples (out of a total of 110) of the use of this adverb in the 'injured leg' sense (example 1) – but use with a speech-act verb (example 2) is overwhelmingly the norm. Example 1 is therefore possible but abnormal. Example 2 is normal.

1. The old dog ambled **lamely** towards them.
2. She hesitated, and then said **lamely**, 'That is all.'

It seems likely that the expected primary sense of this adverb is blocked by the existence of a lexicalization of 'walk lamely', namely *limp*. However this may be, the implications of this tiny example are far-reaching, as they suggest (among other things) that it is more important for lexicographers to research and describe the conventions associated with each lexical item individually than to accept unchallenged the assumptions inherited from theoretical linguists. It is on many thousands of such examples of conventional vs abnormal use of lexical items that the Theory of Norms and Exploitations (TNE; Hanks, forthcoming) is based.

Among many other innovations, Cobuild paid more attention than pre-corpus dictionaries to lexicographic issues such as the role of function words and the pragmatics of discourse organizers (*however, anyway*), which are discussed further in Section 13.6.

At Cobuild, the corpus was used by the lexicographers: (1) to structure the entries, placing the most important meaning of each word first; (2) to write accurate definitions reflecting actual usage; (3) as a source for example sentences and (4) to help decide what to leave out.

The first major impact of corpora on lexicography was therefore on a dictionary for foreign learners with a strong focus on use as an encoding tool. Subsequent newly compiled learners' dictionaries – the *Cambridge International Dictionary of English* (CIDE, first edition 1995), and the *Macmillan English Dictionary for Advanced Learners* (MEDAL, first edition 2002) were also corpus-based and used corpus-derived examples.

In due course complete recensions of the leading English dictionaries for foreign learners, OALD and LDOCE, were prepared on the basis of corpus evidence, though for marketing reasons the distinction between a dictionary as an encoding aid and as a decoding aid tended to be fudged by the publishers and hence by the lexicographers. So OALD (5th edition 1995, 6th edition 2000) is strong on collocates and verb patterns, but the examples are not as natural as either Cobuild or LDOCE, because many of them are deliberately concocted to illustrate underlying patterns of the kind that the editor of the first edition, A. S. Hornby, had described from the outset, rather than being selected without alteration from actual texts. See Section 13.5.

13.3 Coverage: Deciding What to Leave Out

Publishers, their marketeers and advisers (none of whom are lexicographers) often claim that one of the benefits of using corpus evidence is to enable a dictionary to give better coverage of the lexicon of a language. For example, Professor the Lord Quirk (the grammarian Randolph Quirk), wrote in the preface to the third edition of LDOCE (1995) – an edition which was heavily revised using the newly available evidence of the BNC:

> There are two core features of a dictionary in terms of which its degree of excellence and achievement must be measured: **coverage** and **definition**. . . . The advent of computerized corpora enables us to achieve a greatly enhanced coverage. . . . In consequence of new initiatives in coverage, the new LDOCE is about one fifth larger than its predecessors.

Other EFL dictionary publishers have written in similar terms. It is easy publicity to say that the corpus gives better coverage, providing an argument to justify investment in corpora which bean counters can easily understand. However, it is highly questionable whether one-fifth bigger necessarily means one-fifth better. Previous editions of LDOCE already had excellent vocabulary coverage for foreign learners.

The language does not change so fast that thousands of new entries need to be added to a learners' dictionary every few years. A few dozen, maybe, but not thousands. Many new terms are ephemeral and, if added, should be taken out again in the next edition. All too often they are not. The term *black-coated workers* (meaning 'office workers') is a case in point. It was already obsolescent in 1963, when it was added to OALD2, but it took the Oxford lexicographers 20 years to take it out again. So if LDOCE3 is one-fifth better than its predecessors (and in my opinion such a claim would not be unjustified, though hard to quantify), its added excellence is due at least in part to other factors such as sharper definitions and more natural examples, selected from the BNC.

Rather than offering more and more new words to be added, a more valuable benefit of corpus evidence for dictionary compilers – in particular compilers of smaller dictionaries and dictionaries for foreign learners – is that scarcity of corpus evidence can help to give a lexicographer the courage of his or her convictions in deciding what to leave out.

Dictionaries are and always have been full of rare and unusual words. This is especially true of dictionaries intended for decoding use (i.e. those compiled with native speakers in mind, where the lexicographer imagines a scenario in which readers want to find out the meaning of an unusual word more often than they seek information about the correct idiomatic use of more common words). Such dictionaries deliberately err on the side of including rare words because these are the very words that a reader is most likely to look up in the unlikely event of encountering one. However, for compilers of pedagogical dictionaries aimed at foreign learners, this presents an excruciating dilemma. On the one hand, the main purpose of a pedagogical dictionary is to help learners write and speak the language idiomatically. On the other hand, there is nothing more certain to destroy a user's confidence in a dictionary than the experience of looking up a word and not finding it. The problem is compounded when an EFL dictionary tries to meet the needs of learners for both encoding and decoding purposes.

To quote Jonathan Crowther (preface, 1995), the OALD:

> strives to satisfy the . . . basic needs of foreign students . . ., namely to develop their receptive and productive skills, the ability (as Tony Cowie wrote in his preface to the fourth edition) 'to compose as well as to understand'.

This is an ambitious goal, for there is a tremendous tension between the two objectives. Dictionaries such as the *Idiomatic and Syntactic English Dictionary* (ISED) and Cobuild1, which start out with the goal of restricting themselves to the encoding needs of advanced learners, are gradually seduced (mainly for marketing reasons) in successive editions into trying to serve as decoding tools as well, adding thousands and thousands of rare words and senses, which can only be there for decoding users and which clutter up the dictionary for those users who want to use it as an encoding tool.

This desire to meet two objectives is all the more seductive because there is no way of predicting all the many and various needs (both encoding and decoding)

of advanced learners. It is a fair bet that no advanced learner will need to encode a term such as *black-eyed bean* and if one does, it is equally unlikely that he or she will consult MEDAL (where it is an entry) in order to do so. The entry can only be there for decoding users, and it is doubtful whether any of them will go to a learner's dictionary to find out about it.

One of the benefits of using corpus evidence is that it is possible to count the relative frequency of different words and expressions. The second edition of COBUILD introduced a system of diamonds to flag the relative frequency of different words in a corpus. Other EFL dictionaries (e.g. MEDAL, CALD) have followed suit. CALD (the *Cambridge Advanced Learner's Dictionary*) has a 'Frequency Information System' which 'gives students a clear guide to the most important words and meanings to learn':

E – Essential, about 4,900 terms
I – Improver, 3,300 terms
A – Advanced, 3,700 terms

In other words, about 12,000 terms are singled out, on the basis of corpus evidence, as what C.K. Ogden (1930) termed 'basic English' – a reasonable goal for an advanced learner. The blurb of CALD trumpets this feature: 'Frequency information showing you the most important words to learn'.

It must be admitted that dictionaries have not yet been very successful in calculating the relative frequency of meanings as opposed to words. CALD, for example, calculates comparative word frequency with great reliability, and this extends to certain fixed expressions such as phrasal verbs. But it does not do such a good job on the frequencies of different senses of a verb. So, for example, the use of *this* as an adverb, meaning 'as much as shown or to a particular degree' (e.g. *It was only about this high off the ground*) is flagged as E (for 'essential') along with the other uses of this very common word. However, in the BNC, the adverb use accounts for only about 0.16 per cent of all uses of *this*. It occurs 7.47 times per million words in the BNC, being similar in frequency to the words *symmetry*, *fittings*, *specifications*, *trinity* and *bilateral*, suggesting this use is not as essential (in terms of frequency at least) as the authors of CALD indicate.

A short case study of a borderline word will illustrate the encoding/decoding dilemma more clearly. The term *maulstick* is not in my active vocabulary: I had never encountered it and had no idea of its existence, still less about what it meant until, 20 years after first becoming a lexicographer, I stumbled across it in 1985 in the course of doing a lexicographical cross-check. It is in two EFL dictionaries of the 1970s: OALD3 (1974) and LDOCE1 (1978). I hope it is not excessively uncharitable to suppose that the LDOCE1 lexicographers did not dare to leave it out, because their main rival had it in. After all, what other evidence did they have? How could they know, with any confidence, that the word was so rare that it should be omitted? The worst mistake that a beginner in lexicography can make is to omit a term on the grounds, 'I don't know it'. You may not know it, and yet it may nevertheless be a common word or sense for other users of the language.

Lexicographers must make provision for the possibility of their own ignorance. So in it went. Or maybe the editors responsible for the 'maul-' section in these dictionaries were keen amateur oil painters, for, according to LDOCE1, it means 'a stick held by a painter to support the hand which holds the brush'.

By 1995, however, when corpus evidence was brought into play, this entry was deleted from both LDOCE and OALD, and other corpus-based learners' dictionaries have not included it.

That is not the end of the story, however, for there are other kinds of dictionaries besides learners' dictionaries. Even though there is scant corpus evidence for *maulstick* in the BNC (one occurrence), the word receives an entry in the *New Oxford Dictionary of English* (NODE 1998) and its successor, the *Oxford Dictionary of English* (ODE 2003), the only English dictionaries for native speakers that can justify a claim to be corpus-based. Although NODE made very full use of the evidence of the BNC, it did so mainly in order to improve the grammar, definitions and examples for everyday words. Improvements in vocabulary coverage of rarer terms owed more to other sources, in particular the Oxford Reading Programme (in which citation readers search texts for unusual terms and senses) and directed research in particular domains such as flora and fauna, and the vocabulary of sports and leisure activities, and of academic disciplines ranging from nuclear physics to art history. A balanced corpus of 100 million words is not nearly large enough to serve, unsupported, as a basis for an entry list for a native-speaker dictionary. Using corpus evidence to achieve coverage for such a dictionary is necessary, but not sufficient.

In the Oxford English Corpus of 1.5 billion words, there are two hits for *maulstick*. This tantalizing slither of evidence might have presented the lexicographical team with a dilemma – to include or not to include – if they had not already conducted directed research that tells them that it is an important term in art history – and, no doubt, would have substantially greater frequency in a dedicated corpus of art-history texts. Domain-specific corpora have much to contribute to future lexicography, but then questions will arise concerning the overlap between lexicography and terminology, the latter being a pitifully neglected subject in the English-speaking world, though strong in other languages.

In some aspects, dictionaries are being overtaken by other kinds of on-line resources. Anyone who really wants to know what a maulstick is might be well advised to go online (at the time of writing–May 2008, Google has 8,230 citations for the word), while at the relevant entry in Wikipedia, there are two nice pictures (at the time of writing) showing an artist using a maulstick. This is one of those entries where the best explanation is afforded by a visual representation of an example. The grammar, phraseology and definition of *maulstick* are of little linguistic interest.

13.4 Definitions

Does corpus evidence enable lexicographers to sharpen up definitions? It would be nice to be able to answer with an unequivocal 'yes', but the truth is more

complex. The first editions of corpus-based dictionaries sometimes show the lexicographers seeming to struggle to find words to represent what they can see in the corpus, while partly or wholly rejecting definitions inherited from pre-corpus dictionaries. The lexicographers are trying to relate their definitions more closely to how the language is actually used. This calls for a great deal of effort and compression. The effort is sometimes (but not always) successful.

Consider, for example, the adjective *threadbare*. Here is a traditional pre-corpus definition (OALD2/3; 1963, 1974), buried in the entry for *thread*, for which a swung dash is substituted:

> **thread** . . .
>
> . . . **~bare**, *adj.* **1.** (of cloth) worn thin; shabby: *a ~bare coat.* **2.** (fig.) much used and therefore uninteresting or valueless; hackneyed: *~bare jokes (sermons, arguments).*

The apparatus and nesting seem a bit cumbersome to modern eyes, but this is a perfectly serviceable definition. Corpus lexicographers, however, want to get away from the Leibnizian brackets '(of cloth)', which are intended to make the definiens substitutable for the definiendum. They may also think that the corpus evidence does not chime well with 'uninteresting'.

Cobuild1 (1987) has:

> **threadbare**, *adj.* **1 Threadbare** clothes, carpets, and other pieces of cloth are old and have been used so much that the cloth has become very thin: EG *O'Shea's suit was baggy and threadbare.*
>
> **2. Threadbare** jokes, stories, excuses, etc. have been said so often that they are no longer funny, interesting, or believable.

Here, we can see the lexicographer struggling (and failing) to find a suitable superordinate word for 'jokes, stories, excuses, . . .' and eventually giving up, and, in defiance of the editor-in-chief's interdict, employing the forbidden escape word, 'etc.'. Something similar is happening in the second part of the explanation – the definiendum – with 'no longer funny, interesting, or believable.' It seems that what she wanted to say was something like, 'A threadbare joke is no longer funny; a threadbare story is no longer interesting; a threadbare excuse is no longer believable; . . . etc.'

Getting the right level of generalization in lexicography is extremely difficult. Struggling and failing to do so is one reason why many entries in corpus-based dictionaries tend to be rather wordy. Interestingly, it is often easier to find the right superordinate terms in definitions based on samples from very large corpora (billions of words) than in smaller corpora of only one or two hundred million words. The conventional norms of usage for a word tend to stand out in large samples, and these can sometimes suggest an appropriate superordinate. Nevertheless, this is an area in which serious lexicographical training is needed.

Cobuild2 (1995) tries again:

> **threadbare**, *adj.* **1 Threadbare** clothes, carpets, and other pieces of cloth look old, dull, and very thin, because they have been worn or used so much that the cloth has become very thin: *She sat cross-legged on a threadbare square of carpet.*
>
> **2.** If you describe an activity, an idea, or an argument as **threadbare**, you mean that it is very weak, or inadequate, or old and no longer funny, interesting, or believable. . . . *the government's threadbare domestic policies.*

This is hardly more successful. In fact, if anything, it is worse. The words 'dull' in sense 1 and 'activity' in sense 2 are very debatable, while in sense 2 the problem of getting the right level of generalization has not been addressed.

In MEDAL (2003), this is one of the entries which borrows Cobuild's full-sentence style, for two of its three definitions:

> **threadbare**, *adj.* **1** threadbare clothing, carpet, or cloth is very thin and almost has holes in it because it has been worn or used a lot. **1 a.** wearing or containing threadbare things: *the threadbare family apartment.* 2. a threadbare idea or excuse has been used a lot and is no longer effective.

Here at last the problem of the superordinate in sense 2 has been addressed, as indeed it has been in CALD ('a threadbare excuse, argument, or idea . . .'). But is it any better than Hornby's 1963 definition, quoted above? What the corpus dictionaries add – or try to add – is information about the semantic types of collocates: clothing and carpets, not just cloth, in sense 1, and an idea or excuse in sense 2. MEDAL also offers information about a sense extension (1a), typical of many adjectives.

In the first edition of the first corpus-driven dictionary (Cobuild), the definitions are undoubtedly verbose in places, not sharp. Reviewers of the first edition of Cobuild (1987) accused it, with some justice, of verbosity. However, some reviewers went on to associate this with the 'full-sentence' defining style. I think this criticism is mistaken. Let us look at another example. Definition 7 of *proportion* in Cobuild1 reads as follows:

> If you say that something is big or small **in proportion to** something else, you mean that it is big or small when you compare it with the other thing or measure it against the other thing.

This is undeniably verbose. In the second edition, it was reduced to:

> If something is small or large **in proportion to** something else, it is small or large when compared with that thing.

This is a full-sentence definition, but not especially verbose, nor is it significantly longer than the definitions of this difficult concept in other dictionaries.

MEDAL defines **proportion** (sense 2) as:

> the correct, most useful, or most attractive relationship between two things,

and offers the phrase *in proportion to* with an example (*'his head is large in proportion to his small frame'*) but no definition. An undefined example may be the best strategy for such a phrase.

One noticeable effect of corpus evidence is that corpus-based dictionaries – even dictionaries based on different corpora – are tending to converge in what they say about the meanings of words, compared with pre-corpus dictionaries, as described for example by Atkins and Levin (1991), who showed that there was simply no way of mapping the sense distinctions in one pre-corpus dictionary onto another. Such dictionaries were incommensurable. Now, it is clear that there are many different possible ways of carving up corpus data, but corpus-based dictionaries are, in many of their entries, commensurable. They may make more or less fine-grained sense distinctions, but the semantic space being described in two such dictionaries is very often recognizably similar. This is not because they copy from one another or because they are using the same corpus, but because the salient features of word meanings are generally the same across many different corpora. Minor details differ; old decaying senses are more fully represented in some dictionaries than in others, but the salient features of the architecture of a word's meaning are waiting there, to be discovered through painstaking corpus analysis. Corpus lexicography is very often a voyage towards the painful rediscovery of the obvious. After hours of painstaking corpus analysis and hunting for just the right generalizations to cover the bulk of the evidence, you know that you have got it right when your publisher says to you, 'That's obvious. Everyone in the whole world knows that.' To which the corpus lexicographer is minded to retort, 'If everyone in the whole world knows that, why didn't our pre-corpus dictionaries say so?'

A question that engages much lexicographical energy is, 'How many senses are there of this or that word?' To which the riposte is, 'How long is a piece of string?' that is there is no reliable way of deciding how many senses a word has: deciding this is, in each case, a matter of lexicographical art and judgement. Computational linguists often complain that sense distinctions in EFL and other dictionaries are too fine-grained, and this criticism is not totally ill-founded. Striving for a high level of generality obscures many contextually determined nuances. It is also difficult to get right.

Existing dictionary entries are all meaning-driven. A new kind of dictionary is proposed by Hanks and Pustejovsky (2005), which is pattern-driven. In other words, the lexicographers must first sort the corpus evidence for each word into patterns that are associated with it, and then attach a meaning to the pattern, not the word in isolation. An example of a pattern-dictionary entry is cited in Section 13.10.

13.5 Citing Examples

The selection or concoction of examples of usage to illustrate word senses or grammatical points is a vexed question, debated by two camps, with much misunderstanding on both sides. On one side, the editors of some corpus-based dictionaries have argued that only authentic examples – real sentences and phrases which have been uttered in earnest by real people for some real communicative purpose – are acceptable. Made-up examples are viewed as unreliable because they can trample unwittingly over selectional preferences and other unrecognized grammatical constraints and so mislead the user. On the other side, some pedagogical lexicographers argue that the purpose of an example in a dictionary is to illustrate some aspect of the linguistic competence that a dictionary user aims at, not merely to record a performance, and that therefore examples should be idealizations, based on corpus evidence perhaps, but shorter, neater and better focused than most real uttered sentences, which are full of digressions, loose ends and other imperfections. An obvious compromise would be to seek authentic sentences in the corpus that meet the criteria of the idealizers, but unfortunately suitable candidates are few and far between: for many words, they cannot be found at all.

The first edition of one otherwise excellent pre-corpus dictionary was marred by occasional bizarre invented examples, such as (s.v. *salvage*):

> *'We'll try to salvage your leg,' said the doctor to the trapped man.*

There are several things wrong with this. In the first place, legs are not among the things that are normally salvaged (*ships, possessions,* and *pride* are among the more salient collocates in the direct object slot). Second, there are too many players: either 'the doctor' or 'the trapped man', in an authentic text, would probably have been mentioned before and would therefore be a pronoun here. The inventor of this example is trying to tell a whole story in a single sentence, which, of course, does not happen in real texts.

Determining the 'normal' uses of words turns out to be difficult – indeed, impossible without very large bodies of evidence and a theory of prototypical norms. Corpora occasionally throw up bizarre utterances that are implausible but nonetheless authentic:

> Always vacuum your moose from the snout up, and brush your pheasant with freshly baked bread, torn not sliced.
>
> —Example cited by Judy Kegl (personal communication), from *The Massachusetts Journal of Taxidermy*, c. 1986, in an article cited in a corpus of Associated Press newswire texts.

This example is cited from memory, as I no longer have access to that early corpus. It deviates from normal usage in several ways – for example, the noun *moose* is not a canonical direct object of the verb *vacuum*. It would clearly not be a good

example to put in a dictionary, but unfortunately, being human, lexicographers suffer from a temptation to use such examples because they are 'interesting' or because they illustrate some extreme boundary of possible usage. The purpose of a dictionary example is to illustrate normal usage, not the extreme boundaries of possibility.

Bizarre citations such as these have, indeed, been used as arguments to turn lexicographers and linguists alike away from corpus evidence. The obvious questions to ask are, 'What sort of thing do you normally vacuum in English – or is this verb normally intransitive?' The obvious way of answering is to look at the salient collocates in the direct object slot in a corpus (if there is one), either impressionistically or using a statistical tool such as the Sketch Engine (see Culpeper this volume). It is unlikely that *moose* will be found as a salient direct object of *vacuum* in any corpus. Although authentic empirical evidence is a necessary basis for linguistic analysis, it is not in itself sufficient. In other words, authenticity alone is not enough: evidence of conventionality is also needed.

The two sentences just discussed, one invented and the other authentic, are extreme cases on either side. Other badly chosen or badly invented examples are more quietly misleading. Because we cannot be sure that we know all the constraints that govern the idiomatic uses of a word – and because it is very clear that the 'anything goes' syntactic theories of the 1970s were simply wrong, though their legacy is still with us – it is safer to stick to authentic data, rather than making up examples, and to seek examples that are both typical and ordinary. Current dictionaries, both corpus-based and pre-corpus, contain many examples that are quietly misleading in one way or another.

An additional point may be made here about cognitive salience. The fact that, 20 years after hearing the 'vacuum your moose' example in conversation, I can still remember it suggests that it is somehow salient – cognitively salient, that is. I assume that I am a normal human being, at least in this respect, and that others too will find it memorable. It is memorable because it is unusual. But unusual examples do not belong in dictionaries. I remember few if any of the tens of thousands of more mundane sentences to which I must have been exposed in 1987. A large part of everyday language – the frequently recurring patterns, which we might call 'socially salient' – is conventional and for that very reason unmemorable. This suggests that cognitive salience and social salience are independent (or possibly inverse) variables.

Corpus lexicographers need to resist the temptation to select (and even more so, to invent) bizarre examples, regardless of how interesting they may seem. They should instead choose examples in which all the words are used normally, conventionally, and naturally, without unnecessary digressions or distractions. This is difficult.

13.6 Pragmatics

Section 13.2 contained some examples of sentence adverbs, illustrating the impact of corpora on dictionaries in respect of pragmatics. Lexical pragmatics is a very

broad field, with many different realizations, and it is one where corpus evidence has been particularly beneficial. Pre-corpus dictionaries had got into the habit of trying to word all explanations in terms of substitutable definitions, no matter how absurd the result might seem, but corpus-based dictionaries pay much more attention to the pragmatic functions of certain words and expressions. Conversational pragmatics includes terms such as *I see, you know, right* and *of course.* No lexicographer inspecting a concordance for *see, know, right* and *course* could fail to notice the pragmatic force of these expressions. Thus, Cobuild2 has an entry for *right* in its pragmatic functions (separate from the many truth-functional and other semantic meanings of this word), with 'discourse functions' as a part-of-speech label and the following explanations:

1. You use '**right**' to attract someone's attention or to indicate that you have dealt with one thing, so you can go on to another. *Right, I'll be back in a minute | Wonderful. Right, let's go on to our next caller.*
2. You can use '**right?**' to check whether what you have said is correct. *They have a small plane, right? | So if it's not there, the killer must have it, right?*
3. You can say '**right**' to show that you are listening to what someone is saying and that you accept or understand it. (SPOKEN) '*Your children may well come away speaking with a broad country accent*' - '*Right*' - '*because they're mixing with country children.*'
4. You can say '**right on**' to express your support or approval (INFORMAL, SPOKEN, OLD-FASHIONED) *He suggested that many of the ideas just would not work. But the tenor of his input was 'Right on! Please show us how to make them work.'*
5. If someone says '**right you are**' they are agreeing to do something in a very willing and happy way. (INFORMAL, SPOKEN) '*I want a word with you when you stop.*' - '*Right you are.*'

Cobuild's initiative in this respect has been followed by other EFL dictionaries, though not by the corpus-based editions of OALD, which seem to be rather reluctant to let go of the traditional notion that the purpose of dictionary definitions is to define (not to comment on pragmatics).

One more example of the impact of corpus evidence on the description of conversational pragmatics in dictionaries will have to suffice. At its entry for *really*, the corpus-based LDOCE3 (1994) includes a box containing a graph showing that this word is used about 400 times per million words in written English, but approximately 1,800 times per million (i.e. 4½ times more often) in spoken English. The box goes on to illustrate the many different pragmatic uses of this word in speech:

4 **really?** a) used to show that you are surprised by what someone has said: *'There are something like 87 McDonalds in Hong Kong.' 'Really?'* b) used in conversation to show that you are listening to or interested in what the other person is saying. *'I think we might go to see the Grand Canyon in June.' 'Really?'* c) *AmE* used

to express agreement: *'It's a pain having to get here so early.' 'Yeah, really!'* d) especially BrE used to express disapproval: *Really, Larry, you might have told me!*

5 not really used to say 'no' or 'not completely': Do you want to come along? – 'Not really.'

6 I don't really know used to say that you are not certain about something: *I don't really know what he's up to. I haven't heard from him for ages.*

7 really and truly used to emphasize a statement or opinion: *really and truly, I think you should tell him.*

Many other examples of the impact of corpus evidence on the accounting for other kinds of pragmatic information in dictionaries could be given, but lack of space forbids.

13.7 Phraseology

Another aspect of the impact of corpora on dictionaries lies in the area of phraseology. This ranges from highlighting important phrases in examples to providing explicit lists of frequent collocations. Highlighting is a technique favoured by ODE, as in the following example sentence (s.v. *jaw*):

victory was snatched ***from the jaws of*** defeat

and at *plot*, contrasting with the transitive use in 'plotting a bombing campaign', an intransitive example with a salient preposition:

brother plots ***against*** *brother.*

OALD6 contains some useful 'help notes', which address phraseology, among other things. For example, at the entry for *really*, sense 4 ('(usually *spoken*) used, often in negative sentences, to reduce the force of sth. you are saying'), there is a help note that reads as follows:

The position of **really** can change the meaning of the sentence. **I don't really know** means that you are not sure about something. **I really don't know** emphasizes that you do not know.

Many modern dictionaries of current English for learners show lexical selections involving salient collocates based on statistical analysis of corpus data. For example MEDAL at *comfort* adds a note:

Words frequently used with comfort
verbs: bring, derive, draw, find, offer, seek, take

CALD at *threat* has:

> *Words that go with threat*
> be/pose a threat; issue/make a threat; receive a threat; carry out a threat; a threat hangs over sb; a growing/major/real/serious threat; an idle/immediate/potential/renewed threat; a threat to sb/sth; the threat of sth

Information of this kind, which can be extremely useful for encoding purposes, is comparatively easy to select from a corpus, given a good statistical analyser, but would be impossible to dream up out of one's head without corpus evidence. Oxford University Press devotes an entire volume, a companion volume to OALD called the *Oxford Collocations Dictionary for Students of English* (2002), to providing such information for a wide range of common words. The aim is to help learners to enrich their vocabulary and to select idiomatically correct (not merely logically correct) phraseology.

On the other hand, lexicographers must be alert and pay attention to the facts when encoding phraseological information. Several learners' dictionaries include the expression *black economy* in a form that implies that it is always used with the definite article *the*. This would be a useful piece of encoding information for a foreign learner, if it were true. Unfortunately it is not. Out of 43 genuine hits for *black economy* in the BNC, 39 are in a noun phrase governed by the definite article, but 6 are not. This sort of evidence presents lexicographers with another familiar dilemma: whether to represent the rule or to represent the predominant norm.

13.8 Grammar

Corpus evidence has enabled lexicographers to give better, streamlined accounts of English grammar. A notable example is ODE, which, unlike other dictionaries aimed at native speakers, has broken away from traditional simplistic obsessions (in particular, the subcategorization of verbs into merely transitive and intransitive, with occasional mention of prepositional choices). ODE gives an empirically sounder account, based on corpus evidence, of the syntactic patterns associated with each word. For example, ODE recognizes that a verb can have up to three arguments or valencies, and it says what they are: with verbs of movement, adverbial arguments ['with adverbial of direction]'; with linking verbs, a subject complement or object complement (as in 'she dyed her hair black'); and so on.

Old habits of caution in lexicography die hard, however. For example, CALD's entry for the verb **amble** gives the grammar as [I, usually + adv or prep]. 'I' stands for intransitive. In this example, 'usually' is unnecessary. *Amble* is indeed a manner-of-motion verb, so one would expect the adverbial of direction to be optional, but in fact it is obligatory in normal, non-contrastive text. The LDOCE3 entry for this verb is preferable:

> **amble** *v.* [I always + adv/prep] to walk in a slow relaxed way: [**+ along/across etc**] *the old man came out and ambled over for a chat.*

The relevant word in the example sentence here is 'over'. In the front matter, LDOCE3 comments: 'You cannot simply say "he ambled" without adding something like "along" or "towards me".' One needs corpus evidence to be able to make assertions like this in entry after entry with confidence.

Learner's dictionaries have built much of their grammatical apparatus on the insights of pre-corpus lexicographers such as A. S. Hornby, whose verb patterns are justly famous. Hornby and his mentor, H. E. Palmer, perceived the order underlying the apparent chaos of verb use in English. However, it is not surprising that many of the details of Hornby's verb patterns had to be revised in the light of corpus evidence in the 5th and 6th editions of OALD, for Hornby was reliant for the details on his intuitions and his wide reading. He did not have a corpus.

In one recent learners' dictionary, MEDAL, however, the grammatical apparatus is minimal. For example, at *amble*, MEDAL does not mention the more-or-less obligatory adverbial of direction, represented in LDOCE3 as '+ adv/prep'. This is surely a deliberate policy, since the principals involved in creating MEDAL had worked in one capacity or another on other learners' dictionaries, which have more sophisticated grammar patterns. Presumably, it was decided as a matter of policy that MEDAL should focus on meanings and examples, not on grammatical abstractions. After all, many learners learn by analogy, not by rule, so the grammatical abstractions will mean little or nothing to many readers.

13.9 The Role of Corpus Evidence in Dictionaries for Native Speakers

The one-volume *New Oxford Dictionary of English* (NODE 1998; subsequently rechristened the *Oxford Dictionary of English*, *ODE* 2001) is, so far, the only dictionary aimed at native speakers to have made extensive use of corpus evidence to compile a brand-new account of contemporary English for use by native speakers. It made use of three kinds of evidence: the BNC as a template for both the macrostructure and the microstructure of the dictionary and its entries; the Oxford Reading Program for rare and unusual words and senses; and technical literature for information about terminology in special domains, ranging from science to sport and from law to linguistics. *ODE* contrasts with the *Oxford English Dictionary* (Murray 1878–1928; 3rd edition in progress), which is a dictionary compiled on historical principles, placing the oldest meaning of a word first.

To date, *ODE* is the only dictionary of English for native speakers to be corpus-based. The *New Oxford American Dictionary* is an Americanization of it, not an original compilation. In other languages, the situation is rather different – for example, major corpus-based dictionaries for native speakers of languages as diverse as Danish, Modern Greek, and Malay are in compilation or have been published.

The impact of corpus data on lexicography since 1987 (the date of publication of Cobuild, the first corpus-driven dictionary) has been overwhelming. At last lexicographers have sufficient evidence to make the generalizations that they need to make with reasonable confidence. We can now see that pre-corpus lexicography was little more than a series of stabs in the dark, often driven by historical rather than synchronic motives. In word after word, pre-corpus lexicographers

(consulting their intuitions and a bundle of more or less unusual citations collected by worthy and earnest citation readers) failed to achieve the right level of generalization regarding the conventions of present-day word meaning in a language, as can be seen by attempts to map the old definitions onto the new evidence. Of all the many possible uses and meanings that a word might have, lexicographers now have better chances of selecting the ones that are actually used and of writing reasonably accurate descriptive definitions of commonly used words. This has resulted in a clutch of completely new corpus-based dictionaries (Cobuild, CIDE, MEDAL, ODE) as well as completely rewritten editions of old favourites (OALD, LDOCE). But in truth the process of responding to the challenges posed by corpus evidence has hardly begun. What is now called for is a radical reappraisal of lexicological theory in the light of corpus evidence, with close attention to syntagmatics (the way words are normally and actually used), as well as what they mean. This will, if undertaken seriously and objectively, lead to completely new kinds of lexical resources, in particular hierarchically structured multipurpose on-line ontologies and lexicons.

13.10 The Future: FrameNet and the Pattern Dictionary

It seems certain that lexicography in future will be corpus-based, or even corpus-driven. More attention will be paid to the typical phraseology associated with each meaning of each word. Links will be set up between corpus evidence and meanings. One project that is doing this is FrameNet (Baker et al. 2003), which groups words of similar meaning into semantic frames and identifies the frame elements that participate in each frame. For example, in the 'Damaging' frame, there is an *Agent* (the person or thing that causes the damage), a *Patient* (the person or thing that suffers the damage), and a *Cause* (the event that causes the damage). Lexical units identified so far (May 2008) as participants in the 'Damaging' frame include: *chip*.v, *damage*.v, *deface*.v, *dent*.v, *key*.v, *mar*.v, *nick*.v, *rend*.v, *rip*.v, *sabotage*.v, *score*.v, *scratch*.v, *tear*.v, *vandalise*.v, *vandalism*.n. The semantic frame offers additional information about frame elements and lexical units. Annotated corpus examples are given.

FrameNet concentrates on words with similar meaning. A rather different project is the *Pattern Dictionary* (Hanks, in progress), which concentrates on meaning differences and how they can be recognized in texts. The project design is described in Hanks and Pustejovsky (2005); it is currently being implemented as part of the Corpus Pattern Analysis[1] project at the Masaryk University in Brno. The *Pattern Dictionary* aims to account for all the normal uses of all normal verbs in English. This is to say, it is a semantically motivated account of each verb's syntagmatic preferences, providing links between contexts and meanings, that is, there are 'pointers' from each pattern to the uses in a corpus that support it. The *Pattern Dictionary* is not aimed at everyday readers or learners, but is intended as a fundamental resource or benchmark for linguists, lexicographers, coursebook writers, computational linguists and lexicological theorists, with many possible applications. For example, if the Semantic Web (Feigenbaum et al. 2007) gets beyond

processing lists of names and addresses, tagged documents and other entities, and starts to process unstructured texts, it will, sooner or later, have to address the question of what words mean – and how do we know what they mean? *The Pattern Dictionary* provides explicit links between meaning and use. Thus, while FrameNet annotates *scratch* as one of several words participating in the 'Damaging' frame, the Pattern Dictionary distinguishes 14 patterns for the verb *scratch*, only 3 of which have anything to do with 'Damaging'. Several of the meanings of *scratch* participate in frames that have not yet been compiled in FrameNet. Conversely, many verbs that are lexical units in FrameNet do not yet have a *Pattern Dictionary* entry. The two projects are complementary. Some of the *Pattern Dictionary*'s distinctions are quite fine-grained, but they are of vital importance in answering the question 'Who did what to whom?' No distinction is made between semantic and pragmatic implicatures, for both are part of the conventional meaning of these patterns.

scratch

1. PATTERN: [[Human | Physical Object 1]] scratch [[Physical Object 2]]
   PRIMARY IMPLICATURE: [[Human | Physical Object 1]] marks and/or damages the surface of [[Physical Object 2]]
   SECONDARY IMPLICATURE: Typically, if subject is [[Human]], [[Human]] does this by dragging a fingernail or other pointed object across the surface of [[Physical Object 2]]
   EXAMPLES: *I remember my diamond ring scratching the table.* | *'I'm sorry sir, but I'm afraid I've scratched your car a bit!'*
   FREQUENCY: 19%

2. PATTERN: [[Human]] scratch [[Language | Picture]] {on [[Inanimate = Surface]]}
   PRIMARY IMPLICATURE: [[Human]] writes or marks [[Language | Picture]] on [[Inanimate = Surface]] using a sharp edge or other sharp or pointed object
   EXAMPLES: *A Turkish schoolboy who had scratched the word 'Marxism' on his desk.* | *Names of infant Mulverins had recently been scratched on the wall.*
   FREQUENCY: 9%

3. PATTERN: [[Human | Animal]] scratch [[Self | Body Part]]
   PRIMARY IMPLICATURE: [[Human | Animal]] repeatedly drags one or more of his or her fingernails rapidly across [[Body Part]]
   SECONDARY IMPLICATURE: typically, [[Human | Animal]] does this in order to relieve itching
   EXAMPLE: *Without claws it is impossible for any cat to scratch itself efficiently.*
   FREQUENCY: 16%

4. PATTERN: [[Human]] scratch {head}
   PRIMARY IMPLICATURE: [[Human]] rubs his or her {head} with his or her fingernail(s)
   SECONDARY IMPLICATURE: often a sign that [[Human]] is puzzled or bewildered
   EXAMPLES: *He peered down at me and scratched his head as he replaced his cap.* | *Having just struggled through a copy of the Maastricht Treaty I can only scratch my head that anyone would wish to sign it* [METAPHORICAL EXPLOITATION].
   FREQUENCY: 14%

5. PATTERN: [[Human 1 | Animal 1]] scratch [[Human 2 | Animal 2]]
   PRIMARY IMPLICATURE: [[Human 1 | Animal 1]] uses the fingernails or claws to inflict injury on [[Human 2 | Animal 2]]
   EXAMPLE: *Mary was starting to pull her sister's hair violently and scratch her face in anger.*
   FREQUENCY: 9%

6. PATTERN: [[Inanimate]] scratch [[Human | Animal]]
   PRIMARY IMPLICATURE: [[Inanimate]] accidentally inflicts a superficial wound on [[Human | Animal]]
   EXAMPLE: *A nice old Burmese woman brought us limes – her old arms scratched by the thorns.*
   FREQUENCY: 2%

7. PATTERN: [[Bird = Poultry]] scratch [NO OBJ] (around)
   PRIMARY IMPLICATURE: [[Bird = Poultry]] drags its claws over the surface of the ground in quick, repeated movements
   SECONDARY IMPLICATURE: typically, [[Bird = Poultry]] does this as part of searching for seeds or other food.
   EXAMPLE: *A typical garden would contain fruit and vegetables, a few chickens to scratch around.*
   FREQUENCY: 3 %

8. PATTERN: [[Human]] scratch [NO OBJ] {around | about} {for [[Entity = Benefit]]}
   PRIMARY IMPLICATURE [[Human]] tries to obtain [[Entity = Benefit]] in difficult circumstances
   COMMENT: Phrasal verb.
   EXAMPLE: *Worrying his head off, scratching about for the rent.*
   FREQUENCY: 4%

9. PATTERN: [[Human]] scratch {living}
   PRIMARY IMPLICATURE: [[Human]] earns a very poor {living}
   COMMENT: Idiom.

EXAMPLE: *destitute farmers trying to scratch a living from exhausted land.*
FREQUENCY: 6%

10. PATTERN: [[Human 1]] scratch {[[Human 2]]'s {back}}
PRIMARY IMPLICATURE: [[Human 1]] helps [[Human 2]] in some way
SECONDARY IMPLICATURE: usually as part of a reciprocal helping arrangement
COMMENT: Idiom.
EXAMPLE: *Here the guiding motto was: you scratch my back, and I'll scratch yours—a process to which Malinowski usually referred in more dignified language as 'reciprocity' or 'give and take'.*
FREQUENCY: 1%

11. PATTERN: [[Human | Institution]] scratch {surface (of [[Abstract = Topic]])}
PRIMARY IMPLICATURE: [[Human | Institution]] pays only very superficial attention to [[Abstract = Topic]]
COMMENT: Idiom.
EXAMPLE: *As a means of helping Africa's debt burden, . . . it barely scratches the surface of the problem.*
FREQUENCY: 11%

12. PATTERN: [[Human 1]] scratch [[Entity]]
PRIMARY IMPLICATURE: [[Human 1]] looks below the obvious superficial appearance of something . . .
SECONDARY IMPLICATURE: . . . and finds that the reality is very different from the appearance.
COMMENT: Imperative. Idiom.
EXAMPLE: *Scratch any of us and you will find a small child.*
FREQUENCY: 2%

13. PATTERN: [[Human | Physical Object 1 | Process]] scratch [[Physical Object 2 | Stuff]] {away | off}
PRIMARY IMPLICATURE: [[Human | Physical Object 1 | Process]] removes [[Physical Object 2 | Stuff]] from a surface by scratching it
COMMENT: Phrasal verb.
EXAMPLE: *First he scratched away the plaster, then he tried to pull out the bricks.*
FREQUENCY: 2%

14. PATTERN: [Human]] scratch [[Language | Picture]] {out}
PRIMARY IMPLICATURE: [[Human]] deletes or removes [[Language | Picture]] from a document or picture
COMMENT: Phrasal verb.
EXAMPLE: *Some artists . . . use 'body colour' occasionally, especially solid white to give that additional accent such as highlights and sparkles of light on water which sometimes give the same results as scratching out.*
FREQUENCY: 1%

13.11 Conclusion

It would be no exaggeration to say that corpus evidence has had, is having and will continue to have a revolutionizing effect on lexicography. It has enabled lexicographers to get a new sense of proportion about the relative importance of different words and different meanings of words. It has led to the development of entirely new approaches to the lexicographic description of pragmatics, function words, phraseology and grammar. It has led to a heated and potentially productive debate about the role of example sentences in dictionaries. But corpus lexicography is still in its infancy. Computer programs are already in development to improve the selection of typical collocates of each word and typical examples of use. In future, we may expect development of new kinds of lexicographical work, where the microstructure of each entry is *pattern-driven* rather than meaning-driven. In other words, instead of asking, 'How many meanings does this word have, and how shall I define them?' the lexicographer will start by asking, 'How is this word used, how can I group the uses into patterns, and what is the meaning of each pattern?'

Dictionaries Cited

CALD: Woodford, K. et al. (2005), *Cambridge Advanced Learner's Dictionary* (= 2nd edn of CIDE). Cambridge: Cambridge University Press.

CIDE1: Procter, P. et al. (1995), *Cambridge International Dictionary of English*. Cambridge: Cambridge University Press.

COBUILD1: Sinclair, J. M., Hanks, P. et al. (1987), *Collins Cobuild English Language Dictionary*. London and Glasgow: HarperCollins.

COBUILD2, 3: Sinclair, J. M., Fox, G., Francis, G. et al. (1995, 2nd edn; 2001, 3rd edn). *Collins Cobuild English Language Dictionary*. London and Glasgow: HarperCollins.

COD1: Fowler, H. W. and Fowler, F. G. (1911), *Concise Oxford Dictionary of Current English*. Oxford: Oxford University Press.

COD8: Allen, R. et al. (1990), *Concise Oxford Dictionary of Current English*, 8th edn. Oxford: Oxford University Press.

CPA: Hanks, P. (in progress), *Corpus Pattern Analysis: The Pattern Dictionary*. Brno: Faculty of Informatics, Masaryk University: http://nlp.fi.muni.cz/projects/cpa/.

ISED: Hornby, A. S. Gatenby, E. V. and Wakefield, H. (1942), *Idiomatic and Syntactic English Dictionary*. Tokyo: Kaitakusha. Reprinted in 1948 from photographic plates by Oxford University Press as *A Learner's Dictionary of Current English*. See OALD.

LDOCE1: Procter, P. et al. (1978), *Longman Dictionary of Contemporary English*. Harlow: Longman.

LDOCE3: Rundell, M. et al. (1995), *Longman Dictionary of Contemporary English*, 3rd edn. Harlow: Longman.

LDOCE4: Bullon, S. et al. (2003), *Longman Dictionary of Contemporary English*, 'new edition'. Harlow: Longman.

MEDAL: M. Rundell. (2002), *Macmillan English Dictionary for Advanced Learners*. Oxford: Macmillan.

NODE (ODE): Hanks, P., Pearsall, J. et al. (1998), *New Oxford Dictionary of English*. Oxford: Oxford University Press. (2nd edn, 2003 published as *Oxford Dictionary of English*.)

OALD2: Hornby, A. S. et al. (1962), *Oxford Advanced Learner's Dictionary of Current English*, 2nd edn. Oxford: Oxford University Press. (2nd edn of ISED).

OALD3: Hornby, A. S., Cowie, A. et al. (1974), *Oxford Advanced Learner's Dictionary of Current English*, 3rd edn. Oxford: Oxford University Press.

OALD4: Cowie, A. et al. (1974), *Oxford Advanced Learner's Dictionary of Current English*, 4th edn. Oxford: Oxford University Press.

OALD5: Crowther, J. et al. (1995), *Oxford Advanced Learner's Dictionary of Current English*, 5th edn. Oxford: Oxford University Press.

OALD6: Wehmeier, S. et al. (2000), *Oxford Advanced Learner's Dictionary of Current English*, 6th edn. Oxford: Oxford University Press.