The corpus revolution revisited

MICHAEL RUNDELL

An update on the rise and rise of electronic language corpora and their impact on dictionaries

Excerpt from Michael Rundell and Penny Stock, *The Corpus revolution* (*ET*30, 1992)

WHAT Eric Partridge once described as the 'gentle art' of lexicography is going through some exciting and unnerving changes.

Nowhere have these changes been more dramatic than in the area of lexicographic evidence. The past decade has witnessed a revolution in the nature of the linguistic data available to writers of dictionaries, and this has led to the questioning – and in some cases the abandonment – of working practices established over 200 years ago. Underlying all these changes, of course, are the dramatic advances in the capacity of computers to store, access and process text. And, as a result of these developments, the traditional source of evidence for language in use, the lexicographer's citation bank, has now been supplemented by a powerful new resource: the computerized corpus providing concordanced samples of words in context.

Creating and using corpora

The physical gathering of texts for a computer corpus is by no means the simple straightforward task that it might appear from the size of the corpora currently being gathered. It is costly, as texts with paper below a certain quality still have to be keyed, since they cannot be scanned, and getting permission to use texts from copyright owners can be a lengthy and tedious task. Nonetheless, it may seem unnecessary in the corpus age for lexicographers to continue with the pre-technological activity of hand-gathering citations.

The chief advantages of corpus data to the lexicographer lie in the sheer spread and volume of data on a given word and the concentration on the most usual, the most frequent, the most typical. (These advantages are, of course, dependent upon the corpus itself being large enough and sufficiently well-balanced to be reliable.) This means that the evidence tends to concentrate on the core of the language, giving lexicographers excellent evidence for the behaviour of common words, and on the most typical usages of words, giving evidence for collocational and syntactic patterning. A collection of individually searched-for citations tends to be weak in precisely this area. It is astonishingly difficult for even the most experienced person to collect material for ordinary, everyday usages, since human beings tend to notice the unusual. Murray pointed out that in the *OED* collection of citations about fifty instances of *abusion* were sent in as against five instances of *abuse*.

While technology can deliver bigger and better corpora, corpus evidence is in the end just

MICHAEL RUNDELL is still a lexicographer, but the corpora he works with are about 100 times larger than the ones in the 1990s. He is now Editor-in-Chief of learners' dictionaries for Macmillan (including the 'Macmillan English Dictionary, New Edition' 2007) and a Director of Lexicography MasterClass Ltd (www.lexmasterclass.com), a company that runs training courses in lexicography and lexical computing. He is also co-author (with Sue Atkins) of the 'Oxford Guide to Practical Lexicography' (2008).

DOI: 10.1017/S0266078408000060

that: evidence. And the proper function of language experts – be they lexicographers, grammarians, or linguists of whatever variety – is to *interpret* that evidence, to select and synthesize what is significant and appropriate, and so to mediate between the corpus and the end-user of the materials they produce. In this sense, the changes we are witnessing are evolutionary rather than revolutionary. □

I RECENTLY built a corpus of about 2 million words in the field of 'sustainable transport'. The whole process took less than 15 minutes (by which time my new corpus was loaded into a corpus-querying system, ready for use), and it cost me nothing. All which gives some idea of how dramatically the world has changed since Penny Stock and I wrote about the 'Corpus Revolution' in 1992.

At the time, it was not hard to predict that computer processing power and storage capacity would carry on doubling each year. It was already clear, too, that the arrival of the corpus would revolutionize the work of dictionary-makers – hence the title of our articles. These changes were well under way in 1992 and, sixteen years on, their effects are still being felt. In the process, dictionaries have got dramatically better – if by 'better' we mean that the description of language they provide corresponds more closely to the way people actually use words when they communicate with one another. But what we didn't know then was that new technologies were emerging which would have huge implications for lexicographers and linguists (and everyone else). We were not alone in failing to predict the really big change which has transformed the way users access corpus data, and made 'instant corpora' a reality: the arrival of the Web. This is what Donald Rumsfeld would call an 'unknown unknown': the Web is one of those phenomena that few people could even conceive of until it appeared. Though technically 'invented' in the late 1980s, the Web didn't go public till 1993, and it was only in the first years of the twenty-first century – when fast broadband links became widespread – that its full potential began to be realized. The Web has brought far-reaching changes in most aspects of our lives – travel, shopping, banking and keeping in touch with friends, to name a few – and for many people it has become the primary source of information of all kinds. It has also been the catalyst for a second Corpus Revolution.

At the time when Penny Stock and I were planning our series of articles, we were part of the team working to create the British National Corpus (BNC). Back then, developing a corpus was a major (and costly) enterprise, and the BNC – with a goal of 100 million words – was an ambitious undertaking. It took three years to complete and (to quote its own website) 'was a joint effort of a large number of participants, organizations and individuals'. These organizations included three big publishing companies, two university departments and the British Library. In just one corner of the project, a substantial team worked on scanning hundreds of books to convert them to digital form, while numerous others were involved in seeking permissions from copyright holders. The BNC set high standards, both in the quality of its design and in the care with which the texts were processed to yield maximum benefit for users. It also raised the bar in terms of size, dwarfing other corpora currently in use. There was much debate around that time about size (how many words in your corpus?) and content (what kinds of texts it included, and in what proportions?), and about the relationship between these two parameters. We will see later how changes in technology altered the terms of this debate. But first, let us go back and sketch out the background against which we first talked about a 'Corpus Revolution'.

Computers and corpora: 1992 and after

The year we wrote those original articles was also the year when Windows 3.1 was launched. This was a breakthrough moment for the 'graphical user interface' – the now-familiar way that we interact with our computers, but at that time a fairly new, user-friendly alternative to the clunky MS-DOS style operating systems. The arrival of Windows and similar systems marked an important stage in the transformation of the computer from specialized scientific apparatus, operated and understood only by the technically minded, to mass-market consumer product that anyone could use. Even so, relatively few people owned personal computers in 1992. Those of

us lucky enough to be working with corpora back then (still a minority among the lexicographic community) typically accessed our data on expensive desktop machines owned by our employers. (Working from home, as many of us now do, simply wasn't an option, with Internet connections and emailing still in their infancy.) We used a program which generated KWIC (key-word-in-context) concordances by searching the corpus data stored on the computer's hard drive, and several examples of these concordances can be seen in the 1992 *ET* articles.

Some of the things we predicted back then have turned out very much as expected. We talked about 'the new discipline of corpus lexicography', which had begun with monolingual English learner's dictionaries developed in the UK. (The first fully corpus-based English dictionary had been published by the *COBUILD* team in 1987.) That 'new discipline' is now the norm, and it would be odd to find a major dictionary project anywhere in the world which did not take a corpus as its starting point. At the end of 2007, for example, OUP's South African company published a new bilingual dictionary of English and Northern Sotho: the project was based on an analysis of large corpora of both these languages.

Meanwhile, improvements in technology have paved the way for more sophisticated ways of searching the corpus. Suppose, for example, we are writing an entry for the word *sound*. We know this word can function as a verb, as a noun, or as an adjective, and it makes sense to analyse the data for these three wordclasses separately. Suppose, furthermore, that we notice the recurrence of sentences like these (where *sound* is followed by an adjective, optionally modified by an adverb):

She sounded pretty confident
I don't mean to sound ungrateful
At the risk of sounding churlish …
He sounds rather pompous on the phone

It would be helpful if we could examine all such uses by running a specialized corpus search to find every instance of the pattern:

noun/pronoun + *sound* (verb) (+adverb) +adjective

A search like this will only work if the corpus itself is *lemmatized* and *part-of-speech tagged*. A lemmatizer is a tool which automatically relates every form of a word to its 'lemma' (or base form), so that *sounding* and *sounded* are linked to the lemma *sound*. And in a part-of-speech-tagged corpus, all the forms of *sound* as a verb are treated as one lemma, and distinguished from the noun and adjective lemmas. When we enrich the data in a corpus by adding linguistic information of this type, all kinds of queries become possible because our search software now 'knows' when *sounds* is the plural of the noun or when it is the third person singular of the present-tense verb. We already knew about these technologies in 1992, but they had not yet been widely implemented. When we showed corpus data for *staple* and *staples* (in *ET*31), the concordances were still of the more 'primitive' type, showing word-forms rather than lemmas. Thus the concordance for *staple* had noun and adjective uses mingled together:

Pictures of crime and accident victims were a ***staple*** *of the tabloids*
… price controls on ***staple*** *products like bread …*

Had our corpus included a sentence such as 'She now *staples* the two sides together', this would have appeared in the concordance for *staples*. Nowadays, the kind of annotation described above is pretty standard, so we no longer have to trawl through an undifferentiated concordance of *take* or *save* in order to track down noun uses of the former or prepositional uses of the latter.

The arrival of the Web – and a new corpus revolution

In the last of our original articles, we discussed the issues of size and content: what is the optimum extent of a corpus for lexicography, and how can it be 'representative' of the language of which it is a sample? The need for large volumes of data was explained in terms of the 'Zipfian' distribution of vocabulary: in most languages, there is a small number of very frequent words, and what we would now call a 'long tail' of many infrequent words. These principles have not altered, and the arguments for collecting large corpora remain compelling. What has changed, though, is that data scarcity is no longer an issue for most languages. The arrival of the Internet, and its extraordinary growth, has put at our disposal more or less infinite quantities of digitized text in a wide range of registers, and this has become the raw material for contemporary corpora (in English

and many other languages). Oxford University Press, for example, now has a huge and diverse English corpus made up entirely of texts from the Web. The 'Oxford English Corpus' (OEC) already stands at over 2 *billion* words – an order of magnitude bigger than the 'conventionally' gathered corpora of the 1990s. To quote its own website (http://www.askoxford.com/oec/), it 'represents all types of English, from literary novels and specialist journals to everyday newspapers and magazines and from *Hansard* to the language of chatrooms, emails and weblogs'. Back in 1992, we could not have foreseen how this new data-source would change the way corpora are created, nor the way it would stimulate the development of new kinds of corpus-querying software.

The downside of having very large amounts of linguistic data is that the task of analysing it becomes a lot more difficult. Even in 1992 we were starting to wonder how we would cope with much larger corpora. It is arduous enough to scan a concordance of three or four hundred lines – but what happens when a corpus search returns 10,000 hits for a word or pattern we are interested in? Fortunately, new software tools have come to the rescue. Concordances are now complemented by what we call 'lexical profiles'. A lexical profile is an automated summary, which illustrates how your search-word behaves in all its main grammatical relationships. A well-known type of lexical profile is the 'Word Sketch', a feature of the Sketch Engine corpus-querying package (http://www.sketchengine.co.uk). Figure 1 shows part of a Word Sketch for the noun *impression*. Among other things, it tells us that *impression* frequently appears as the object of a verb, and we get a list of the verbs that most regularly occur in this structure. Profiles like this provide lexicographers with a revealing overview of a word's main characteristics. They tell us not only which other words regularly occur with our search-word, but also in which types of syntactic pattern it is normally used. And if a word has a marked tendency to appear in one particular form, the software will tell us: so, for example, the Word Sketch for *arrest* alerts us to the fact that it has a strong preference for being used in the passive. Lexical profiles do not replace 'traditional' concordances, but for many lexicographers they have become the primary tool for analyzing words.

This combination of new search routines and vastly larger corpora has simplified the issues regarding corpus design. In a small corpus, a single 'rogue' text has the potential to skew the data, for example by spuriously inflating the importance of certain lexical items. There were good reasons, therefore, why the 1-million-word Brown Corpus of 1962 was designed with such great care. But, as corpora grow ever larger, and we analyze them mainly through lexical profiles (which focus on frequently recurring behaviour and ignore anything uncommon), we no longer need to select our corpus texts with this degree of precision. In a billion-word corpus, the occasional oddball text will not compromise the overall picture, so we now simply aim to ensure that the major text-types are all well represented in our corpus. The arguments about 'representativeness', in other words, have lost some of their force in the brave new world of mega-corpora.

It is fair to say, then, that the arrival of the Web has sparked a second Corpus Revolution. At the beginning of this article, I mentioned a 2-million-word corpus I built in under fifteen minutes. This was made possible by a software tool that comes with the Sketch Engine. It collects tranches of continuous text from the Web, cleans it up to remove links, images, lists and other Internet garbage, then lemmatizes and part-of-speech tags the resulting data to create a ready-to-use corpus. All of which has, in a sense, democratized access to language data: you don't have to work for a large publishing company to be able to use a corpus. But the Web is not only a source of free, already-digitized data (doing away with the need to scan or key in texts). It is also the channel through which we access the data. For today's lexicographer, the usual working method is to view and analyse corpus data online, so we no longer need to install either the search software or the corpus itself on our own machines.

Some things haven't changed

Despite these technical advances, some things haven't changed. Acquiring high-quality spoken data remains a difficult and expensive business, at least if we want to capture spontaneous face-to-face interactions. The use of citations (short extracts from texts, selected by human readers following a tradition that goes back at least as far as Dr Johnson) still has a place in lexicography, though now mainly as a way of tracing word histories or monitoring

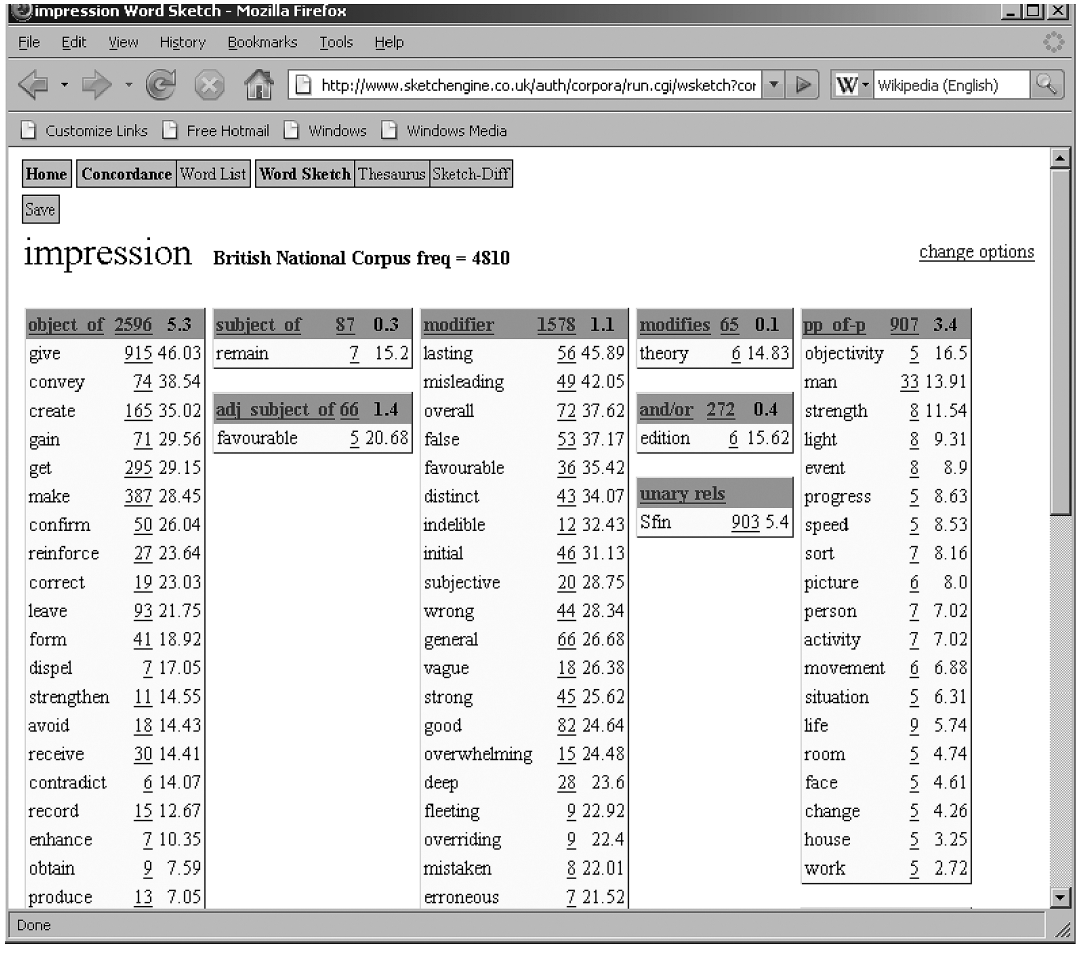

Figure 1 A 'Word Sketch' for *impression*

changes in the language. And the goal of bringing the corpus into the language-teaching classroom (though vigorously pursued by a small group of enthusiasts) remains as elusive as ever. Most startlingly, while lexicographers worldwide now routinely use corpus data, the big US dictionary publishers continue, inexplicably, to behave as if corpora did not exist.

Back in 1992, we pondered the relative importance – in the dictionary-making process – of lexicographers' intuitions about language as opposed to what the empirical evidence tells us. With so much data now at our disposal, and so many more ways of extracting relevant facts from it, the issue has become less critical: where a given usage can be shown to be both frequent and widespread, there is no questioning its status as being 'in the language' and therefore worth recording. Yet good intuition is still a valuable faculty for anyone engaged in describing languages. Though the language resources available to us have improved dramatically, creating high-quality dictionary text from these raw materials remains very much a human skill. As lexicographers, we have the same goals we always had: of producing a description of language which is faithful to the available evidence and well adapted to the needs of the people who will use it. Thanks to technology, and to two corpus revolutions, we are better placed than ever to achieve these goals. ■

Note

This article is dedicated to Penny Stock, who wrote the original 'Corpus Revolution' series with me. Penny sadly died in 2006. The lexicographic community lost an original and insightful mind – and many of us lost a good friend.