DOI 10.1515/phras-2012-0001 — YoP 2012; 3: 1–2

Koenraad Kuiper

Editorial: Corpora and phraseology

This volume of the *Yearbook of Phraseology* like the previous one contains a number of articles which make use of corpora. Corpora consist of text, which some linguists call 'real language', that is they are the result of parole, the output of linguistic performance. They thus provide evidence of the way in which language has been used and raise the general question of what kind of data performance outputs are and what they can tell us.

What is notable about corpora as data sets? First they are just that, data sets, which in themselves tell us nothing until they are interrogated. Second, they are data sets collected for a particular purpose. For example, large balanced corpora may have been assembled to aid lexicographers while corpora of folktales from the Balkans were probably not collected for lexicographers. Third, there are no complete corpora. Because human linguistic output is too large, assumptions have to be made as to how representative a corpus is for the purpose. The corpus of texts in Old English may appear to be complete since there are a relatively small set of extant texts in Old English. But even for such a corpus, there were undoubtedly other texts in Old English which are not extant, not to mention all the talk speakers produced in early medieval England over a period of 400 years.

It is also notable that corpora, by their very nature as text stored on computers, are able to be subjected to certain kinds of computational manipulations but not others. Corpora may be tagged; they may be searched; they may be concorded. The outcome of those processes are lists of items from the corpus, and statistics related to the items in the corpus. Since all phraseological units have collocational properties, phraseologists have a legitimate interest in corpora and what can be found out using them.

There are some caveats. While interrogating a corpus can yield statistical evidence for kinds of lexical cohesion such as one gets with ngrams and lists of above-chance collocates for a head word, interrogating the corpus cannot indicate whether or not these computational artefacts bear any relationship to what is in the native speaker's brain. Suppose that, as a result of hearing and reading a language, an individual has a corpus in the mind. Readers of Shakespeare may have whole plays mentally stored and opera singers, librettos and small children, the admonitions of their parents. How does such a mental corpus relate to a speaker's knowledge of their language and how does it relate to what they do with that knowledge? These are largely open questions. No doubt for the opera singer the memorised libretto enables her to sing a role but does that have any influence on her knowledge of Italian? That is why after the extraction of a computationally derived candidate list so as to obtain recognisable phraseological units, a manual deletion process usually follows.

It is also important to ask what relationship the corpus in the mind, that is the internal corpus of a speaker, and the corpus in a computer have to one another. A large balanced external corpus such as the British National Corpus (BNC) contains texts which no internal mental corpus will contain. So, when a frequency is computed for a linguistic unit from the BNC, what does that say about the frequency of that unit in the internal mental corpus of an individual?

The internal mental corpus is likely to be associated with representations of the contexts in which the texts were acquired. The mental corpus is also parsed and semantically interpreted. However, this is not the case with many external corpora. Many candidate expressions extracted from external corpora are not represented in the mental lexicon since they have no meaning and are syntactically incoherent. For example *and the* is a lexical bundle in terms of its collocational properties and its frequency of occurrence. But is it also a phraseme?

This leads us back to the fundamental questions of what are phraseologists interested in and how corpus data can provide useful evidence to support their enquiries. There are no easy answers to these questions. It does not follow that, just because corpora exist, they will provide useful data for every scholarly enterprise in which phraseologists are engaged. Speakers know lexicalised phrases which are not part of a corpus but part of a lexicon which is extracted somehow from both the corpus of texts a speaker has heard and read, and related to the speaker's own mental corpus. We can be fairly sure that phrasal lexical items are not extracted only on the basis of their frequency since many have a very low frequency, even in very large external corpora. So we need to see corpora as manifestations of a much more complex social order, which we cannot (yet) store alongside the linguistic forms in an external corpus although speakers do store representations of their experiences alongside their internal corpora and lexical items.

This year again, this volume would not have come into existence without the able and co-operative efforts of all the editors and the reviewers. I thank them and the contributors for making volume three possible.