WORDS, CORPUS AND BACK TO WORDS: FROM LANGUAGE TO DISCOURSE

Miguel Fuster Márquez
Moisés Almela

Last century's revolution in computer technologies has also brought with it some changes in the way we conceive language, which are partly due to such revolution, though not entirely. Technological advances in the field of information and communication have made the compilation and processing of large amounts of data an incredibly easy and fast task. Until quite recently, the compilation of large amounts of text was a job that required an enormous effort by researchers. At present, such process has become more feasible and certainly less time consuming, giving the researcher more freedom to think about interesting ways of exploring the data.

However, other important 'revolutions' have taken place in linguistics which in various ways have been favoured by these technological developments. One such important revolution has to do with linguistic theorisation. Linguists in the past would have been happy to decide on language matters simply by asking themselves how the grammar of their mother tongues worked since, as native speakers, they felt to be competent enough to take such decisions. This mentalistic approach, of course we are oversimplifying such approaches considerably, relied on the introspective mental power of well-educated speakers, and for most insightful decisions they made on the matter at hand they did not need to observe the authentic language produced by other speakers. All they needed was their own knowledge and their analytical power. In the

» Fuster Márquez, Miguel & Almela, Moisés. 2017. "Words, Corpus and back to Words: from language to discourse". *Quaderns de Filologia: Estudis Lingüístics* 22: 9-12. doi: 10.7203/qf.22.11297

famous Saussurean dichotomy between ‘langue’ and ‘parole’, these linguists were on the side of ‘langue’; ‘parole’ was of little or no interest. However, an important change that was taking place in linguistics was one in which other linguists started to give priority to the manifestations of ‘parole’; that is, how language was actually used by speakers in their communities in order to theorise with greater accuracy about ‘langue’, or linguistic competence. Various significant developments are related to such more empirical linguistic movement. One of these was the acknowledgement of the spoken language as a legitimate part of language. Twentieth century lexicographers started to collect and introduce examples of informal or conversational registers in the dictionaries they produced. Also, no less significant in this new approach was, for example, the thrust of sociolinguistics, a broad research field, with many branches and fuzzy boundaries, that viewed languages as heterogeneous entities. Sociolinguists observed that variation was more the rule than the exception in speech communities. Sociolinguists brought with them empirical methodologies that enabled them to analyse how real speakers produced language in real settings in order to build their theories of variation and change. Sociolinguistics also made use of quantification in their methodologies. This is partly the context for the emergence of corpus linguistics as a new approach to language. The new framework relied on the examination of real data that had its origin in language use, to build convincing linguistic arguments. Both variation and usage have been essential arguments in corpus approaches.

However, a corpus should not be confused with a database, quoting Sinclair (1996: 2.1) “[a] corpus is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language.” In contrast with any collection of data – any corpus linguist would insist – a corpus contains a representative sample of language if the researcher needs to draw relevant conclusions about language. Broadly speaking, unlike essentially mentalistic approaches, corpus research is empirical, with a preference for inductiveness, that is, the careful analysis of data in representative corpora.

However, most practitioners would agree that corpus linguistics is not a theory, it is a methodology, even if such a methodology is somehow special. In fact, such methodology may be applied to a language, different languages, different varieties of language or registers, by means of small, medium or large corpora, and adopt different approaches in order to test different theories. Interest in corpus linguistics today may refer to areas such as the quality of corpus compilation, lexis and phraseology, grammar, variation and change, discourse or stylistics, among others. Corpus linguistics has been of interest in theoretical and applied linguistics. There is abundant applied research, for example, in the fields of lexicography, second language acquisition or translation. Indeed, it is difficult to think of research areas where corpus linguistics does not have room and something important to offer.

Quite regularly, corpus methodology combines quantitative and qualitative approaches; where, in fact, one approach feeds the other. Former purely qualitative analyses have been in many cases superseded by approaches where quantification and statistics are becoming more prominent. Nevertheless, many convinced corpus linguists would also claim that they are in favour of triangulation and convergent evidence as a more acceptable approach.

Very frequently, the procedure of a corpus linguist will have as its starting point a word or a word list. Therefore, the close examination of a word's behaviour will be crucial for practically any kind of research which relies on language use. It is also known that the most significant advances in contemporary lexicography have been driven by the inspection of reference corpora of variable size and scope that have allowed researchers a more thorough understanding of real usage. Also, the compilation of comparable corpora has provided the basis for establishing parallels, differences and nuances for the purpose of comparability or contrast between languages. In addition, the possibility of compiling more specialized ad hoc corpora has allowed the detailed analysis of vocabulary in different types of discourse, either to determine its value in specialized languages or to gain a better understanding of social or ideological implications, which is determined by the evaluation of linguistic preferences. Finally, it should be added that corpus approaches have revealed the existence of linguistic units which go beyond more traditional lexicological approaches. Extensive research on phraseology and corpus-based lexicography produced in recent decades has brought to light the frequency in discourse of meaningful co-occurring lexical patterns and lexical-grammatical co-selection.

The aim of this issue is to bring together investigation into the lexicon in a variety of languages, in a diversity of manifestations – both at the word level and beyond the word level – and from a variety of perspectives, including not only those which focus on how the vocabulary is internally organized, but also those which deal with the role that lexical units and lexical relations play in the organization of other language levels, particularly in the organization of discourse. These issues are approached from a variety of perspectives that include not only developments in several disciplines of theoretical and descriptive linguistics, particularly in lexicology, phraseology, word formation, discourse analysis, but also in diverse applied disciplines such as translation, foreign language teaching, English for specific purposes and critical discourse analysis. One of the criteria employed in the compilation of the volume was also the coverage of linguistic diversity. In total, six different languages are investigated in the studies selected in this volume: English, German, Spanish, French, Portuguese, Italian. Without claiming exhaustiveness, we consider that the variety of contributions presented here offers an insight into the vigour of current corpus research into phenomena related to the lexicon. Admittedly, the full range of topics, approaches and methodologies developed in this area of research could not fit in a single volume, but a careful selection of studies representing a variety of interesting advances can be representative of significant developments taking place in the field.

References

Sinclair, John McH. 1996. *EAGLES. Preliminary Recommendations on Corpus Typology*. http://www.ilc.pi.cnr.it/EAGLES96/corpustyp/corpustyp.html.