Corpus linguistics, grammar and theory: Report on a panel discussion at the 24th ICAME conference

Jan Aarts

University of Nijmegen

1. Introduction

At the 24th ICAME conference, a panel discussion was held on grammar and linguistic theory in relation to corpus linguistics. The panel members were: Bas Aarts, Geoffrey Leech, Christian Mair, Joybrato Mukherjee and Elena Tognini Bonelli. The panel was chaired by Jan Aarts. Since the title of the panel was rather broad, it was decided to focus the discussion somewhat by concentrating on the question of what impact the wide availability of ever-larger corpora might or should have on the long tradition of English reference grammars, a tradition that goes back to the beginning of the twentieth century and that has seen, in the last few decades, the publication of two grammars originating from the Survey of English Usage (Quirk et al., 1972 and 1985) and, in recent years, the addition of two new grammars: Biber et al. (1999) and Huddleston and Pullum (2002).

The panel opened with short ‘position statements’ from the panel members in which they put forward what each of them thought needed to be said most about the future of reference grammars. The discussion that followed and in which the audience took a large active part concentrated mainly on two issues: the characteristics (both actual and desired) of a reference grammar, and the corpus-linguistic methodology appropriate for the writing of such grammars as well as for corpus-linguistic research in general. The issue here was mainly the choice between ‘corpus-based’ and ‘corpus-driven’ research, where the corpus-based approach uses corpus data as a testbed for pre-conceived hypotheses, while the corpus-driven method puts observation first and wants to arrive at a theory accounting for the data inductively. The latter approach was recently put forward most cogently by Tognini Bonelli (2001).

In the present report, the position statements given by the panel members are presented first; they were written by the panel members themselves. The report of the discussion that followed is presented thematically, following the two major issues in the discussion mentioned above. The names of the participants in the discussion are not mentioned, although there are some verbatim passages. It has been attempted to stay as closely as possible to what people actually said.
2. Position statements

2.1 Bas Aarts (University College London):

I regard myself first and foremost as a syntactician, more specifically a syntactician interested in English. In researching English syntax, I make use of any useful resources that come to hand. To find out about previous thinking on any particular topic I use books, articles, conference papers and the like, found in the traditional way by browsing in libraries and trawling through bibliographies, as well as by using web resources like OPACs, bibliographical databases and the like. For my data I use introspection as well as corpora. In short, in everything I do I’m what you might call an *instrumentalist*, by which I mean that I will make use of any methodological approach that delivers the goods, and gets the job done.

As far as corpora are concerned, I believe that they represent invaluable resources for finding data, and I have used them in all my work. My definition of a corpus is wide one: it encompasses any collection of texts, both written and spoken, without necessarily being marked up in any way. A pile of *Guardian* newspapers from, say, the year 2000, would constitute a corpus. I do not therefore agree with Danny [Joybrato] Mukherjee when in a review of the recent *Cambridge grammar of the English language* on the LINGUIST list he writes:

> [T]he Wall Street Journal, in my view, does not qualify as a representative ‘corpus’ but is an example of a linguistically unstructured ‘archive’ (which may be used as a source of authentic examples but from which general trends in language cannot be extrapolated). (LINGUIST List 13.1853)

While it is true that trends in language cannot be extrapolated from a corpus as used in the sense that I specified, there is nevertheless no reason not to use the term for my pile of newspapers. If someone uses a corpus (in my sense) simply to find data for the particular point of grammar they are investigating, then the issue of tracking trends in language simply does not arise. If a researcher wishes to be more precise about what type of corpus he or she has used for their research, they can say that they used a ‘parsed corpus’ or a ‘tagged corpus’, or a ‘parsed and tagged corpus’ or even simply ‘an annotated corpus’. Danny says:

> Can a reference grammar of the English language, published in the year 2002, really be based on corpus material containing three million words only? I would say no. Although information obtained from corpus-based dictionaries and grammars has been taken into consideration in various regards (e.g. for lists of words that tend to be frequently used in a grammatical construction), the use of corpus data remains unsystematic because there is no discussion of how the data