“In studying corpora we observe a stream of creative energy that is awesome in its wide applicability, its subtlety and its flexibility.”


“I CAME, but where are we going?”

(Title provided by Jeffrely Leech for the Panel of ICAME-25)

Introduction

The Conference that took place at the University of Verona on 19-23 May 2004 was intended to mark the silver jubilee of the meetings and conferences of the International Archive of Modern and Medieval English (ICAME) Association. As such, it attracted a number of scholars from all over the world, who contributed over a hundred paper presentations, thus leading the organisers to set up four parallel sessions, which then materialised as two separate books of proceedings; the first one – co-edited by the present author together with Matti Rissanen – was published early this year under the title Corpus-based Studies of Diachronic English (Bern, Peter Lang), as indeed it focuses on diachronic studies, while the present volume is related to synchronic research and aims to provide a fairly broad and thematic overview of the work undertaken in the field of computerised corpus linguistic studies from their origin to the present day, without overlooking their future prospects. The papers are grouped under three separate headings, which will be illustrated in detail in the following sections.

Section 1: Overviewing twenty-five years of corpus linguistic studies

Looking back “to the stone age of corpus linguistics” (p. 12), the time of language corpora B.C., that is Before Computers, Jan Svartvik opens the lively carousel of detailed reports delivered by the plenary speakers of the conference, by recalling the massive corpus-based work carried out by James Murray, Alexander Ellis, Otto Jesperson, and Charles Fries. When the era of computerised corpora started, from the sixties onwards, two linguistic ‘households’ appeared and confronted each other: ‘armchair linguists’ and ‘corpus linguists’. With great verve, Svartvik illustrates these two eras, how he has lived through them, how he has decided to become and to remain a corpus linguist, and finally how he has contributed to the realisation of the first family of corpora – the Brown corpus and its cognates.

Antoinette Renouf complements Jan Svartvik’s paper perfectly and widens its breadth, by providing a plethora of information and data concerning corpus development as it used to be in the early years of corpus linguistics, as it is nowadays and as it might be in the future. With detailed exemplifications, she touches on a number of corpora, some of which she has helped create herself: the
Birmingham Corpus, the Bank of English, the British National Corpus, open-ended corpora of journalism, and even the World Wide Web used as a corpus. She discusses the major motivating sources underpinning the development of these and other corpora and touches on theoretical issues like corpus ‘representativeness’ and on terminological aspects like the difference between ‘corpus’, ‘database’ and ‘text archive’ or again between ‘historical’ and ‘diachronic’.

In “Seeing through multilingual corpora”, Stig Johansson opens up a new folder in the corpus linguistics archive and testifies to the keyness of multilingual/parallel corpora in linguistic research; indeed, by covering two or more languages, they enable us to see more clearly not only the characteristics of each language, but also the features shared by individual languages and, maybe, even what typifies Language in general. While illustrating the typicality of these corpora, Johansson also supplies readers with essential theoretical and terminological concepts, including the definition of ‘parallel’ and ‘translation’ corpus, the notions of ‘translation paradigm’, of ‘mutual correspondence’ between particular forms or categories of forms, of ‘zero correspondence’, in the case of no formal cross-linguistic correspondence, and finally of ‘parallel translation’. These concepts are exemplified mostly from the English-Norwegian Parallel Corpus, from the Oslo Multilingual Corpus, including German, English and Norwegian, and from the English-Swedish Parallel Corpus, all of the three owing a lot to Stig Johansson’s enlightened ideas.

Finally, from written to spoken: Anne Wichmann devotes her contribution to “Corpora and spoken discourse”, as the title reads; with richness of detail and well-grounded argumentation, she ponders the potential and limitations of computerised corpora of spoken English from their early stages to the present. She starts from the London-Lund Corpus and the Spoken English Corpus, the only two prosodically transcribed, then focuses on the Cobuild Corpus and on the British National Corpus, and finally deals with the specialised Reading Emotion Corpus and the Oxford/Cambridge IViE Corpus, whose software provides annotation tiers linked to the speech signal. Her overview ends up with the Aix-Marsec Corpus: its compilers exploited a dictionary to capture phonemic representation automatically. While discussing the different ways of identifying features of speech, Wichmann convincingly argues that transcription details may provide a unique source of contextual information and are thus important, if not mandatory, for proper corpus analysis, in terms of tagging, parsing, and demographic/situational details. Yet she also warns that the analysis of transcriptions should not mislead the user into thinking the spoken recordings themselves are no longer relevant.

Section 2: Descriptive studies in English syntax and semantics

The description of language and the exploration of its structures, with special perspective on the interconnections between syntax and semantics, have always been among the privileged areas of study in corpus linguistics, if not the most privileged. Indeed, it is now widely acknowledged that grammar and lexis,