Downsizing and upgrading: Why we need more spoken, more multilingual and more nonstandard corpora

Christian Mair (University of Freiburg, Germany)

Today, students of English (and a few other mostly European languages) are privileged in that they can rely on extremely rich corpus-linguistic working environments. In a brief review of 50 years' corpus-linguistic research I will demonstrate how the availability of increasingly large corpora and increasingly sophisticated tools for analysis has left a profound mark on the discipline of linguistics. Traditional descriptive work can now be carried out to higher empirical standards. More importantly, new areas of linguistic inquiry have been opened up to rigorous empirical investigation, and corpus-based research has given a general boost to usage-based theoretical frameworks of all kinds.

As I will show, however, the story of the past fifty years has not been one of undiluted progress and success. It seems that a "conspiracy" of technological and ideological factors has favoured the creation of large monolingual standard written corpora. Data which does not fit this template tends to be made to conform to it. For example, much corpus-based work on spoken English is based on transcriptions rather than the original audio or audiovisual recordings. Similarly, complex multilingual realities tend to be simplified in corpus-compilation, for example by annotating code-switches into other languages as "extra-corpus material."

Today, corpus technology and corpus-linguistic theorising have advanced to such an extent that these biases can and should be redressed. In the digital textual universe in which the humanities and social sciences are all operating today, the classic definition of the corpus, as a usually digital database compiled by linguists for the purposes of linguistic analysis, has become increasingly difficult to uphold and corpus-linguistics will sooner or later merge with the digital humanities movement. A kind of corpus-linguistics which emphasises spoken, multilingual and nonstandard data more than has been the case in the past will make a richer contribution to this development.