Introduction: new frontiers of corpus research

Pam Peters, Peter Collins† and Adam Smith

Macquarie University †University of New South Wales

This volume presents a selection of papers from the 21st ICAME conference (International Computer Archive of Modern and Medieval English), which was held over Easter 2000 at Macquarie University in Sydney, Australia. It was the first ICAME conference to be held in the southern hemisphere, and designed to provide fresh perspectives on corpus linguistics through their application in places bordering on the Indian and Pacific oceans. This put a focus on Asian speech communities who use English as a second language, as well as antipodeans whose usage both reflects and varies the norms of native-speaker Englishes in the northern half of the world. The conference featured several papers on the discourse of younger users of English, and on its form and use in communicating via electronic and telephonic media, as well as print. The conference agenda also included papers from ICAME’s established areas of strength, with new research into English grammar, discourse and the lexicon.

All these frontiers are documented in the present volume. Fresh correlations between genre and language variation are identified. Diachronic change and English language history are illuminated with data from standard and newly created corpora. The methodology of corpus creation, and tools for interrogating and manipulating corpus data, are features of various papers below. Some of the greatest technical challenges come from the data contained in historical corpora, though the computational techniques developed there have application to modern non-standardized English.

The three sections of this volume group together papers with broadly similar focuses:

- new corpora and new speech communities
- historical and regional studies
- corpus-based language description

Such headings are not of course discrete, and in reviewing the papers from each section we will note comparable themes and issues in the other sections – to provide the print equivalent of hypertext links for the interested reader. For ease of reference, the papers in each section are ordered alphabetically by their authors’ surnames.

1. New corpora and new speech communities

Classic standard corpora such as the Brown, LOB and BNC are of course intended to represent a broad spectrum of usage (within the broadly defined region), rather than narrowly defined communities of users. Their existence allows a second generation of corpus builders to concentrate on specifiable groups of English-users, not necessarily native speakers, and on the functions of English discourse for them.
One such corpus is profiled in McEnery’s paper on the 100 Corpus (a collection of telephone calls to the operator for information and services). The highly goal-oriented nature of the exchanges makes for lexical convergence and a limited range of politeness strategies. Brekke’s paper also focuses on a corpus of relatively constrained discourse, comprising economic-administrative texts. His particular interest is in developing a term bank of English and Norwegian economic terminology for Norwegian business. The research will help to achieve a realistic balance between the pressure to use English business terms, and the desire to avoid losing those available in Norwegian. Both McEnery’s and Brekke’s papers highlight actual or potential links with industry for corpus research.

Other papers in the first section feature special-purpose corpora of new English-speaking communities, especially in Asia. The usage of Chinese learners is captured in the papers by Drave (from Hong Kong) and He Anping (from Guangzhou), in aspects of spoken and written discourse. Drave’s paper describes the elements of vague language found in the English of Chinese and native-speakers e.g. sort of, and so on, and about with a precise number. He Anping shows that the techniques of ‘vaguification’ are quickly taken up by language learners, although the repertoire of Chinese speakers of English is smaller and less varied than that of native English speakers. By contrast, He’s paper shows that the particle so is overextended in the writing of Chinese students of English, probably because it parallels the Chinese particle ‘gum’ which has wide-ranging discourse functions. The interactive discourse of Singaporean students (mostly of Chinese origin) produces a few local idioms, in Ooi’s analysis of a corpus of internet relay chat (IRC). But most of the IRC lexis involves truncated English words and phrases found in global netspeak (e.g. pls ‘please’, gd ‘good’, lol ‘laugh out loud’, gtg ‘got to go’).

The standard corpora of English have tended to embody the usage of middle-aged to older writers, because of their dependence on published material. The usage of younger writers and speakers of English has remained largely hidden, and so papers focusing on their discourse are a welcome development. Ooi’s paper is one such, as is Minugh’s with its profile of students’ written language from a corpus of college newspapers in the USA, Canada, the UK, Ireland, Australia, South Africa and New Zealand. Minugh’s analysis shows both their conforming side and some of their more outrageous detail in the use of slang such as shag, ass, techno, fave. In between those extremes, we note that these younger writers make more use than their elders (in data from Brown, Frown, LOB and FLOB) of such things as albeit and sneaked rather than sneaked. Generational differences like this are the stuff of language change, as demonstrated by Lehmann [section 2].

Compiling corpora of spoken language is notoriously labor-intensive, yet eminently rewarding for researchers if strategically planned. The construction of the corpus of spoken Dutch is detailed by Oostdijk in the final paper of this section. The project is a major step towards describing the varieties of Dutch spoken in the Netherlands and Flanders (though not South Africa), and in many contexts, public and private. Oostdijk’s work, like that of McEnery and Brekke, has strong links with European industry.