Pam Peters, Peter Collins and **Adam Smith** (eds.). *New frontiers of corpus research. Papers from the Twenty-First International Conference on English Language Research on Computerized Corpora, Sydney 2000* (Language and Computers series 36). Amsterdam and New York: Rodopi, 2002. 333 pp. ISBN 90-420-1237-4 (bound). Reviewed by **Bernard De Clerck**, University of Ghent. Fund for scientific research – Flanders (Belgium).

Introduction

The volume contains a selection of papers presented at 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 is divided into three major sections, which group together papers with broadly similar focuses. Section I ("New corpora and new speech communities") contains seven papers, Section II ("Historical and regional studies") contains six papers, while Section III ("Corpus-based language description") contains eight papers. Additionally, the volume contains a short introduction by the editors, a general index and a list of contributors. Papers within each section are ordered alphabetically by their authors' surnames.

Summary and discussion

As the different sections suggest, the volume contains a wide spectrum of papers ranging from instances of more traditional, descriptive and comparative corpus research, based on classical corpora (LOB, FLOB, Brown, Frown, BNC) to work-in-progress reports on newly created corpora and the technical challenges they present to existing corpus linguistics (and related software programs). Especially the papers of the latter kind (grouped under section I) live up to the expectations raised by the title of this volume. They all reflect research in new (kinds of) corpora and address the problems these kinds of data present. Of course, this does not mean that papers under sections II and III, while not exactly opening up new frontiers, are inherently less interesting. In fact, they embody traditional instances of ICAME's established areas of strength, with new and valuable research and insights into English grammar, discourse and lexicon. By means of an ingenious use of and comparison between standard and ad hoc corpora, they present valuable findings and emphasise the importance of corpus linguistics in the process of raising and answering questions about the structure of discourse, changes in the nature of English grammar and its lexicon,

and about the communities of people who use the language for diverse purposes. What follows is a brief summary and discussion per section.

New corpora and new speech communities

Nearly all papers in this section exemplify a trend in corpus linguistics showing a growing interest in the construction and analysis of corpora that do not, unlike the classic standard corpora such as Brown, LOB and BNC, intend to represent a broad spectrum of usage. The second generation of corpus builders seems to concentrate on those domains of language usage that have been somewhat underrepresented in the traditional corpora (e.g. language usage by young people, or other varieties of English apart from AE and BE). This is partly caused by the fact that the compilation of corpora by the first generation, which originally aimed at capturing standard language, heavily relied and depended on written published material. This resulted in a somewhat biased representation of language usage: a large amount of the data represents standard written usage by middle-aged to older native speakers. The corpora presented in section one have the potential to fill up some of the gaps. Papers by Drave and He Anping, for example, feature corpora of new English-speaking Asian communities (Chinese EFL-learners) and compare the use of vague language found in the English of Chinese and native speakers. Needless to say, this kind of data-driven research provides a source for more accurate teaching materials and deeper insights into the nature of intercultural miscommunication.

The corpora presented in the papers by Minugh and Ooi kill two birds with one stone: on the one hand they cover the usage of young writers and speakers of English (who, apart from the COLT corpus, have not received ample attention) and on the other hand they capture these data in the hybrid forms of speech and writing stemming from the "new" electronic communication media (basically every possible means of communication by means of a computer, ranging from e-mail and SMS to IRC) that have become ever so manifestly present nowadays. Minugh's COLL-corpus consists of web-based college students, newspapers in the USA, Canada, the UK, Ireland, Australia, South Africa and New Zealand. The data shows a mixture of formal language (things such as *albeit* and *snuck* instead of *sneaked*) and informal language (slang) in a way serving their own communicative purposes and reshaping the language that is handed over to them. Once this corpus is completed, it will indeed provide a very useful source for synchronic and diachronic research that aims at unravelling aspects in language change as it manifests itself in the generational differences illustrated by the data. The data presented in Ooi's analysis of a corpus of internet relay chat (IRC) presents a challenge to corpus linguistics. With the exponential growth of

this new type of data, the agenda of corpus linguistics should include refining appropriate computational tools and linguistic theories in order to provide a more thorough investigation of new linguistic patterns and lexis, involving truncated English words and typical phrases found in global netspeak (e.g. *pls* "please", *gd* "good", *lol* "laugh out loud", *gtg* "got to go").

The corpora used in the papers by Brekke and McEnery can be specified as assembled data coming from a certain specialised domain of usage. Brekke's ongoing construction of an English-Norwegian text-and-term database (TERMINEC), for example, focuses on data from relatively constrained discourse, comprising economic-administrative texts. It will form the basis for student-oriented bilingual domain glossaries with definitions, as well as genre-related material for learning and teaching, not to mention the different doors it will open to further research. One of the aims of this research is to achieve a realistic balance between the pressure to use English business terms and the desire to avoid losing those available in Norwegian. McEnery's paper on the 100 Corpus (a collection of telephone calls to the operator for information and services) focuses on lexis, indirectness and politeness in TTDs (Telephone-based transactional dialogues) and highlights the lexical convergence and the limited range of politeness strategies, caused by the highly goal-oriented nature of these exchanges. Research on these domains will throw light on the specific linguistic characteristics of these data and will help to improve communication within these fields.

Nelleke Oostdijk's paper describes the various considerations that have guided the design of the pioneering Spoken Dutch Corpus (Corpus Gesproken Nederlands) as it is currently under construction. The project aims at a large-scale collection and description of the spoken varieties of Dutch in the Netherlands and Flanders in different public and private contexts. The paper once more exemplifies the very laborious and complex nature of corpus compilation.

Historical and regional studies

Papers under this section are marked by an interest in changes in English usage over anything from several decades (e.g. the paper by Lehmann) to a few centuries (e.g. the papers by Hoffmann, Johansson). The papers by Schneider and Markus highlight the technical problems in dealing with historical corpora.

Gotti's paper discusses the use of the Chadwyck Healey CD-ROM of Early English Prose Fiction to identify canting terms and expressions and to verify their meanings in comparison with those given in contemporary lexicographic work. The presence of at least a dozen of these canting terms and expressions in the EEPC proves that they were not restricted to the 17th and 18th century dic-

tionaries and glossaries. The reading of some of the quotations found during these searches has also pointed to the occurrence of terms in earlier sources than those commonly attested. In this way Gotti's paper provides a more accurate description of the topic at hand.

Hoffmann's paper is a corpus-based study (using texts from Project Gothenburg as well as texts from the Zurich English Newspaper corpus (ZEN)) of complex prepositions (such as *in spite of*, *in front of*, *on (the) face of*) over the last three centuries. He traces their frequency, distribution and development and analyses them in terms of grammaticalisation and related changes in meaning in both literary and non-literary sources.

Johansson's study analyses the history of pied piping (that is, the placement of a preposition immediately before a (relative) *wh*-word) and stranding (placement of the preposition at the end of the sentence). The study is based on the Penn-Helsinki Parsed Corpus of Middle English, containing Middle English prose texts from 1150–1500. It appears that stranded prepositions in Middle English are rare, but that the pre-stages of stranding can be discerned, such as constructions with both pied piping and stranding. The period between 1350 and 1420 offers the largest number of instances of prepositional relative constructions, as well as the largest variety of constructional types.

Lehmann's paper comprises a large-scale corpus study focusing on the use of zero relative constructions in spoken American and British English. One of the great merits of this study is that it proves the possibility of retrieving zero-elements from an electronic corpus by means of a tag-based retrieval strategy. The study shows that there is a sharp difference between AE with 2.5 per cent and BE with 13 per cent of subject relatives with zero relativizer. Together with results from historical linguistics, the data lends support to the existence of a still ongoing language change involving the loss of zero subject relativizers in BE (especially among younger users). AE seems to have undergone the same language change, possibly under the influence of immigrant languages that use fully articulated relative constructions.

For his contrastive study on 15th and 17th century English letters, Manfred Markus compiled a subcorpus of 84 letters from the Innsbruck Letter Corpus with equal amounts of male and female correspondence from the 15th and the 17th centuries. The analysis reveals a few characteristic differences between the two centuries, but its main finding is that women generally show more empathy in their letters, using the channel of communication more actively, more intuitively and more co-operatively. It should be remarked, however, that not only the gender of the writer but also that of the intended recipient will probably have

a bearing on the writing itself. This is another interesting aspect that could be further explored.

The paper by Schneider describes the ongoing development of a software spelling normalisation system (ZENSPELL) that is used to assign normalised, present-day English spellings to 18th century spelling variants, while keeping the source text intact and available for comparison. The idea is to create a target text with 18th century sentences, but with 20th century orthographic words. As such, the target text can be used as input for word class taggers and makes it possible to carry out lexical searches by means of one normalised search term. This technique of automatic disambiguation will be of great help to historical corpus linguistics.

Corpus-based language description

The papers included in this section use corpora of spoken and written language or multiple genres (e.g. the different subcorpora in the ICE-GB), or they start from traditional descriptive grammars (or intuitions of contemporary speakers as in De Haan's paper) and use corpus evidence to support, refute or further specify these descriptions or intuitions.

De Haan's paper on the frequency and use of the relative pronoun *whom* is a nice example of how native speakers/linguists' intuitions (which claimed that *whom* has virtually disappeared from spoken language) can be proved wrong or incomplete by means of corpus study. The analysis shows its continuing use in different text categories, especially the more formal types of writing, although it is by no means confined to writing.

Facchinetti's study on the modal meanings of *can* and *could*, based on a ten per cent sample of the ICE-GB, shows the interdependency between differences in use, meaning and text type. There is a discrepancy in both frequency and semantic values between the two modals. The use of "ability" *can* is concentrated in the spoken medium, while epistemic *could* is common in academic writing.

Holmes and Sigley's paper is both descriptive and historical in its use of corpus data to track social changes in patterns of gender marking between 1961 and 1991. Focusing on frequency data for general terms and the use of gender pre- and post-modification, the analysis identifies differences between patterns found in the Brown and LOB corpora, compared to the Frown, FLOB and Wellington corpora. The results suggest that women continue to be the linguistically marked gender but that there is some evidence to support a positive interpretation, since the marked contexts reflect the entry of women into occupational domains that were previously considered as exclusively male.

Kjellmer's paper proves how corpora can be of help in charting the semantic development of certain lexical items. His study of *eventual* shows that the word has two meanings, one associated with finality ("final", "ultimate") and another with potentiality ("possible", "potential"). Although the second meaning is labelled as "archaic" in dictionaries of contemporary English, Kjellmer notices a restoration of this use in spoken BE, creating an open-choice situation between potentiality and finality. This could be caused by the influence of AE or Continental languages and, interestingly, translation practices (and the influence of false friends) at the EU in Brussels.

Ilka Mindt's paper on the patterns of falling intonation, using data from the Lancaster/IBM Spoken English Corpus, documents the typical pattern of declination across individual texts and speakers, but also shows how this can be varied to support the speaker's purpose at particular points in the discourse. The downward trend is analysed as being normal and unmarked, whereas an upward trend of starting points and endpoints across a text is interpreted as conveying paralinguistic meaning.

The paper by Rayson et al. examines the relationship between part-of-speech frequencies and text typology in the BNC Sampler (text types used were informative writing, imaginative writing, conversation and task-oriented speech). The actual variation gradient 1) conversation, 2) imaginative writing, 3) task-oriented speech and 4) informative writing shows that genre and medium interact in a more complex way than was originally hypothesised. This exemplifies the heterogeneous nature of some text type categories in corpus linguistics and calls for further research.

Just like De Haan, Schlüter uses data from corpora to verify intuitive impressions on the use of temporal adverbials in combination with the present perfect. Corpus study shows that only one out of three utterances with the present perfect co-occurs with a temporal adverbial. Furthermore, Schlüter remarks that Alexander's list (1988) does not include the most frequent temporal adverbials occurring in the consulted corpora. Again – as shown in Facchinetti's paper on *can* and *could* – specific distribution varies across different registers.

In his comparative study of the LOB and FLOB corpora, Smith traces the substantial changes and growth of the progressive in BE. The most impressive rise of the progressive occurs in the present tense, where it is realised by a wider range of verb types, and appears increasingly more in main clauses than in subordinate clauses. However, Smith remarks that, while this tendency looks like a "pure" grammatical change, one also has to bear in mind shifts in stylistic change, resulting in different attitudes towards the inclusion of data with ele-

ments of colloquial speech habits in the more recent compilations of written corpora.

Conclusion

One of the major strengths of this volume is that it encourages further research. It does so in two ways. First of all, it highlights the merits of corpus linguistics by presenting papers of traditional corpus research leading to a better description of the language system on a syntactic, semantic and pragmatic level. Secondly, it does so by presenting papers that move beyond existing corpora and tackle intriguing problems in the compilation and analysis of new corpora representing instances of language usage that have been underrepresented so far (new varieties of English and the use of English through new communication media). This new knowledge acquired by corpus linguistics might eventually lead to practical applications in the real-life situations from which the data is taken. This harmonious process of give-and-take may hopefully narrow the gap between the academic and non-academic world. That in itself will definitely open up new frontiers. Looking forward to the sequel.