Bernhard Kettemann and **Georg Marko** (eds.). *Teaching and learning by doing corpus analysis: Proceedings of the Fourth International Conference on Teaching and Language Corpora, Graz 19–24 July, 2000*. Language and Computers: Studies in Practical Linguistics 42. Amsterdam and New York: Rodopi, 2002. 390 pp. ISBN 90-420-1450-4. Reviewed by **Anne Curzan**, University of Michigan.

This volume of papers from TALC-2000 takes as its premise the value of corpora and data-driven learning to effective, student-centered teaching and learning in language classes, be they focused on second- or foreign-language learning, on linguistics, on translation studies, or on teaching for specific or for academic purposes. John Kirk, in his contribution to the volume, notes the larger shift from "teaching" to "learning" in pedagogical theory, and this volume exemplifies this pedagogical approach in practice, using corpora as the means for autonomous student-learning experiences. In the process, the volume both reaffirms the value of incorporating corpora into learning languages – or about languages – and highlights exciting innovations in available or developing corpus-based resources and pedagogical strategies. Many of the papers will be of interest to language and linguistics instructors designing student-centered, corpus-based linguistic investigation, as the authors share their best practices, cautions, successes, and failures in working with corpora. Other papers discuss issues at the heart of corpus design and will appeal to a perhaps even broader audience in corpus linguistics. The papers are both retrospective and forward-looking, as this community of scholars shares their experiences in order to further the development and exploitation of language corpora in teaching, learning, and research.

The volume's twenty-three generally very concise papers are divided into six sections: it begins with a broader section on "General Aspects of Corpus Linguistics" (four papers) and a short section on "Corpus-based Teaching Material" (two papers); at the heart of the volume are eight articles in "Data-driven learning"; the last three sections focus on more specialized corpora and language learning – "Learner Corpora" (three papers), "Corpus Analysis of ESP for Teaching Purposes" (three papers), and "Corpus Analysis and the Teaching of Translation" (three papers). The editors of the volume rightly note that these section headings capture only one way to categorize the papers and that contributions extend across category boundaries. In the brief introduction to the volume, Tony McEnery usefully clusters the articles within the framework of the developing focus of the TALC conferences, particularly on multilingual corpora

and on the increasing variety of corpus-based approaches and applications. In this review, I create a different set of connections, drawing out themes that run through the volume. In a relatively short review of a volume containing this many papers, I want to acknowledge upfront that I cannot do justice to the detailed content of many of the articles. In highlighting this set of issues and questions that run through the volume, I hope to demonstrate the richness of the contents from multiple perspectives.

Several papers describe new projects, often still in progress and/or representing early steps toward more ambitious designs. Federico Zanettin, in "CEXI: Designing an English-Italian Translational Corpus," introduces this bilingual, parallel, bidirectional, translation-driven corpus of English and Italian. Zanettin carefully leads readers through the selection process that resulted in the CEXI corpus of all books that have been translated between 1976-2000, have been published in Italy, the United Kingdom, or the United States, and are directed at adult readers – a process that raises questions of desired representativeness in these kinds of specialized corpora (an issue revisited in several other papers in the volume). Mike Scott provides a summary of the Guardian Keyword Database, to accompany the CD-ROM included with the volume, in "Picturing the Key Words of a very Large Corpus and their Lexical Upshots or Getting at the Guardian's View of the World." Keywords, their associates, and the calculation of "clumping" appear to be a feasible way to process very large databases and view structured hierarchies of "aboutness" or occasionally of stylistic features – although, as Scott notes, this work and its implications are currently exploratory. In "The Influence of External Factors on Learner Performance," Ylva Berglund and Oliver Mason describe their attempt to perform automatic stylistic analyses of texts using low-level features (e.g., sentence length, type/token ratios, average word length) as a way to identify how language-learner data differ from the production of native speakers. Their pilot study compares data from the Uppsala Student English corpus (USE) and from a subset of Frown; it would be very interesting to match these results with a comparison of data from USE to data from a corpus of native-speaker student essays, which may be significantly less polished than the material in Frown.

Agniezka Leko-Szymaska, in "How to Trace the Growth in Learners' Active Vocabulary?: A Corpus-based Study," exploits the PELCRA corpus of learner English, in comparison with the results of Vocabulary Level Tests which demonstrate receptive knowledge, to test the validity of two measures of lexical richness. Her useful conclusion that the Condensed Lexical Frequency Profile is the most meaningful measure of lexical richness is equally valuable for the methodology employed to test the measures and reach this conclusion.

Several papers raise fundamental questions about corpus design, including large representative corpora and smaller, more specialized ones. Lou Burnard's paper, "The BNC: Where did we Go Wrong?," offers a useful, concise retrospective on the development of the BNC, from acquiring permissions to sampling techniques, from annotation and encoding to distribution. Burnard's candor makes this paper an important read for prospective corpus designers. Near the end of the paper, Burnard discusses the repositioning of the BNC as a repository of language diversity (versus a "representative" corpus) and the insufficiency of the taxonomy of text types to exploit the BNC effectively in this regard. David Lee, four sections later, provides an almost direct response in "Genres, Registers, Text Types, Domains and Styles: Clarifying the Concepts and Navigating a Path through the BNC Jungle."

Lee's contribution is in at least two ways anomalous in the collection: at forty-five pages, it is two to three times longer than any other paper; and the first half of the paper is an extended theoretical treatment of an ongoing terminological and conceptual issue in the field that is only indirectly related to the use of corpora in classrooms. That said, it is a very smart, interesting treatment of the distinction among the terms *register*, *genre*, and *text type* that provides a wide-ranging survey of the published material on the topic and stakes a well-justified position on how best to categorize texts in corpora, drawing on insights from prototype theory. In terms of corpus design, Lee argues that we are interested in genres, and he provides one of the clearest, most persuasive distinctions of *register* and *genre* that I have seen published; as he summarizes: "I contend that it is useful to see the two terms *genre* and *register* as really two different angles or points of view, with *register* being used when we are talking about lexicogrammatical and discourse-semantic patterns associated with situations (i.e. linguistic patterns), and *genre* being used when we are talking about memberships of culturally-recognisable categories" (p. 260). (It should be noted that scholars in genre theory would probably push his description of genre further in terms of genre's constitutive power of rhetorical situations.) Drawing particularly on work by Gerard Steen, Lee argues for genre categories at the basic level, where genres are maximally distinct; many existing corpora, he points out in a survey of ICE-GB, LOB, and the BNC, mix supergenres, genres, and subgenres in their "genre" classifications. After an extended critique of the BNC categories and titles, Lee offers the BNC Index (which works from three existing resources), "a comprehensive, user-friendly, 'one-stop' database of information in the BNC" (p. 274). Lee notes that some decisions were, of course, subjective, and some corpus users may disagree with his decisions, but the taxonomy and decisions are laid out plainly here. Importantly, the Index opens the possibility of creating

specialized sub-corpora for research or teaching/learning. And in the first paper of the volume, "The Learner as Corpus Designer," Guy Aston argues persuasively for the benefits of asking students to extract sub-corpora from larger ones, which can be specifically targeted and provide learners experience in corpus design and evaluation.

Laura Gavioli, in "Some Thoughts on the Problem of Representing ESP through Small Corpora," addresses the lack of an agreed-upon set of criteria for adequate representativeness of small corpora. Her description of students' work with small corpora of medical research articles highlights some of the benefits and pitfalls of these small ESP corpora. Claire Kennedy and Tiziana Miceli ("The *CWIC* Project: Developing and Using a Corpus for Intermediate Italian Students") argue for accessibility over representativeness in smaller corpora, given their work teaching Italian to intermediate students in Australia. The Contemporary Written Italian Corpus (CWIC) is made up of interactive, short (whole) texts that can serve as models of expert performances in the types of texts students must themselves produce. This way, students can find models for expressing particular rhetorical moves and answer their own questions along the lines, "Should I use X or Y here?" The rationale behind CWIC echoes Averil Coxhead's argument earlier in the volume: language teachers should teach materials which are directly relevant to the learners ("The Academic Word List: A Corpus-based Word List for Academic Purposes"). Coxhead is primarily focused on the implications of the Academic Word List for vocabulary learning and teaching, but the principle clearly applies to the design of many of the specialized corpora described in the volume.

John Flowerdew, in "Computer-assisted Analysis of Language Learner Diaries: A Qualitative Application of Word Frequency and Concordancing Software," describes perhaps the most specialized corpus in the volume: students' diaries reflecting on their learning a language as preparation for ESL teaching. Flowerdew exploits the corpus only in using keywords to locate stretches of text that capture the learners' preoccupations – it demonstrates the potential of corpus data as a qualitative research tool to assess the effectiveness of a program.

Several papers directly address the ways in which corpus-based learning empowers students. Tim Johns, in "Data-driven Learning: The Perceptual Challenge," describes learners as detectives, who when confronted with data must draw conclusions from clues. The practical examples here, from teaching collocations with prepositions to helping graduate students learn more nuanced collocational patterns, will be of interest to many ESL instructors. In "Exploring New Directions for Discovery Learning", Silvia Bernardini describes learners browsing with teachers as guides and also offers very useful specific examples, such

as teaching adverb and adjective collocations and helping students find patterns that vary by register (the benefit of incorporating multiple corpora). Natalie Kübler points out how much fundamental linguistics students are required to learn – and are motivated to learn – in working with corpora to learn authentic, specialized English. The kind of discovery learning and problem solving required by querying corpora highlight important issues in natural language processing for students and allow them to go beyond dictionaries to examine specialized meanings and syntactic environments. All of these articles describe a close, interactive relationship between instructors and students that defies any notion that computerized learning can (and should?) lead to distance learning – a concern that Christian Mair wisely raises at the end of his paper, given the widespread "technophilia" at universities and the circulation of ideas like "virtual universities."

Offering a different cautionary note about these kinds of corpus-based, discovery-learning experiences in the ESL classroom, David Wible, Feng-yi Chien, Chin-Hwa Kuo and C. C. Wang argue that unfiltered examples, which may surpass the lexical range of less advanced students, can actually be detrimental ("Toward Automating a Personalized Concordancer for Data-Driven Learning: A Lexical Difficulty Filter for Language Learners"). They describe a new tool, the Lexical Difficulty Filter (LDF), which filters examples based on the frequency of the words in the line and can be adjusted to different thresholds of lexical difficulty. The LDF, they assert, assures that concordancing tools are not restricted as "elite" tools and can simulate more specialized corpora by extracting examples from larger corpora.

John Kirk, in a paper focused on teaching linguistics as opposed to ESL ("Teaching Critical Skills in Corpus Linguistics Using the BNC"), stresses the importance of students learning critical skills (in a systematic manner, with systematic assessment) in addition to querying or concordancing skills. For example, through replicating the searches in published corpus studies, students learn to assess others' methodologies as a step toward designing their own studies.

In "Empowering Non-Native Speakers: The Hidden Surplus Value of Corpora in Continental English Departments", Christian Mair focuses specifically on empowering non-native speakers through the use of corpora. Corpora allow students to test the judgments of native speakers and the authoritative prescriptions of grammar books. As Mair points out, "using the appropriate corpora, any student can disprove statements made in the most authoritative reference grammar of English in less than half an hour" (p. 124).

Two papers in the volume demonstrate how corpus-based work can trouble traditional grammar categories. Gunter Lorenz, in "Language Corpora Rock the

Base: On Standard English Grammar, Perfective Aspect and Seemingly Adverse Corpus Evidence", calls for a distinction between perfective aspect and perfect forms, as part of a larger argument for replacing the teaching of "Good English" as the model for teaching ESL with the teaching of a "multi-layered, multi-variety standard of English" (p. 132). "Adverse" corpus findings are, in fact, a critical component of this constructivist approach to grammatical "rules." In "A Corpus-based Grammar for ELT", Dieter Mindt describes his corpus-based grammar of the English verb system (published in 2000), in which the categories are inductive. His five classes of verbs, which categorize *have to* and *like to* as catenative, offer a fascinating new way to think about the distinction between finite and non-finite verb phrases. The examples typically provide frequencies of different constructions and the entire approach of the grammatical descriptions targets English language learners.

Three other studies of verbs round out the volume. Paul Thompson focuses on core modal auxiliary verbs in selected agricultural theses in the Reading Academic Text corpus ("Modal Verbs in Academic Writing"). He concludes that EAP material tends to overemphasize the hedging role of modals; his study indicates that in various rhetorical sections within a thesis, modals serve other important functions, such as objective modality. Noëlle Serpollet tests to see if mandative *should* is decreasing through a comparison of LOB, FLOB and INTERSECT (a bilingual French-English translation corpus) and examines how mandative *should* is translated into French ("Mandative Constructions in English and their Equivalents in French – Applying a Bilingual Approach to the Theory and Practice of Translation"). Normalizing all the frequency counts could have strengthened the case made here, but it does demonstrate how bilingual corpora can aid translation studies and teaching. Claudia Claridge, in "Translating Phrasal Verbs," investigates the translation of selected English phrasal verbs into German using the Chemnitz English-German Translation Corpus. She outlines five different translation strategies evidenced in the corpus, noting the frequency and variation between translations with German particles and prefixes. Bilingual corpora clearly hold exciting possibilities for the teaching of translation and translation studies.

The world-wide web is undoubtedly, perhaps inevitably central to future developments in corpus linguistics, particularly the monitoring of ongoing language change. Antoinette Renouf tackles the question of how to study recent language change in "The Time Dimension in Modern English Corpus Linguistics" – an effective complement to the primarily synchronic focus of the rest of the volume. After describing the journalistic corpus compiled at Liverpool and the software they have developed there, Renouf outlines two systems for man-

aging the web and calls for the development of more resources and methodologies. Although we have yet to figure out how to manage "the diversity and unpredictability" of the web, to quote Burnard (p. 68), it has the potential to allow corpus linguists and their students to keep up with language change during the necessary gaps created by the time-consuming process of corpus compilation, annotation and encoding, and distribution.

The constraints of a volume with this many papers mean that authors often have to summarize and gesture towards the richness of their pedagogical approaches, of their corpora and tools, and of their studies; at the same time, this conciseness allows the volume to capture the wide range of work happening in corpus-based teaching and learning. There are minor glitches in the editing of the volume – for example, Mair's abstract seems to have been replaced by a duplicate of Bernardini's and Lee is left out of the list of contributors – but these do not distract from the content of the volume. I wished that some of the reproduced images could have been clearer, but they remain readable.

The volume as a whole highlights exciting developments in approaches to teaching and learning with corpora and in the development of resources and methodologies relevant to research as well as teaching. It stresses the importance of discovery learning – both in the classroom and in research. As one of Bernardini's students puts it, after working with corpora: "There is little certainty left: relying on intuitions, even regarding one's own native language, becomes more problematic …" (p. 179). The papers in this volume highlight the value of studying spoken and written language in use, captured in modern corpora, in terms of learning language, translating it, and studying it for linguistic description.