Reviews

Douglas Biber, **Susan Conrad**, and **Randi Reppen**. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press, 1998. 300 pp. ISBN 0-521-49957-7. Reviewed by **Anne Curzan**, University of Washington.

In August of 1991, thirty-four scholars gathered in Lidingö, Sweden for a Nobel Symposium on modern corpus linguistics, in many ways marking the field's 'coming of age'. According to the rigorous Nobel criteria, modern corpus linguistics had proven itself to be a field of great scientific importance and great relevance to society (Svartvik 1992: 12). Since then, the publication of corpus-based research articles has continued to grow exponentially, new corpora have been developed, and existing corpora have been expanded. A more recent development in the field has been the publication of corpus linguistic textbooks—yet another sign that the field is becoming ever more firmly established in wider academic circles, as scholars in the field develop the material to educate and train students. The newest textbook contributed to the effort, Biber, Conrad, and Reppen's *Corpus Linguistics: Investigating Language Structure and Use* (hereafter *CL*), with its focus on the new research possibilities opened by modern corpus linguistics, demonstrates in many ways how far the field has come, even within the past decade. As the subtitle suggests, this book is as focused on new insights into the structure and use of language as it is on describing corpus linguistics, the methodology employed to gain this new understanding. It also represents a concerted effort by the authors to present the field in a way highly accessible to students new to linguistics.

Biber, Conrad, and Reppen (hereafter BCR) work from the premise that a corpus-based approach can 'shed light on previously intractable research questions in linguistics' (p ix). Unlike many previous books and articles addressing modern corpus linguistics as a field, BCR's book does not begin with an explicit defense of corpus linguistics in contrast to rationalist linguistics. For decades, Chomsky's unforgiving condemnation of corpus-based approaches to language study has been reverberating in the scholarship; for example, Fillmore (1992) frames his article as an attempt to reconcile the armchair linguist and the computational linguist; Biber and Finegan (1991) begin an earlier article by recounting an exchange between R. B. Lees and W. Nelson Francis on the merits of building a corpus; and McEnery and Wilson (1996) devote much of the first chapter of their textbook to an evaluation of rationalist criticisms.

BCR's book, part of a new generation of corpus linguistic scholarship, no longer explicitly reflects this perceived need to defend its own merits. The authors' rejection of *purely* rationalist approaches, which appears several times in the introduction, is kept general and concise; for example, 'comprehensive studies of use cannot rely on intuition, anecdotal evidence, or small samples; they rather require empirical analysis of large databases of authentic texts, as in the corpus-based approach' (p 9).[1] BCR assume (rather than argue for) a combined empirical and intuitive approach in any corpus-based study; in fact, this joint qualitative-quantitative approach is listed in the introduction as one of four essential characteristics of modern corpus linguistics (p 4). As BCR note, the critical contribution of corpus linguistics to our understanding of language is the study of the actual language used in naturally occurring texts—including variation, frequency, and association patterns—rather than the study of what is theoretically possible in language.

CL is designed as an introductory teaching text, best targeted at undergraduate students. It assumes little to no background in linguistics, and the authors are careful to explain the linguistic terminology that they employ (eg *morphology, lemma*). The questions and explanations are kept basic and clear; the structure of the book involves significant repetition, reminding students of what they have read and how it relates to what they will read. All the example analyses focus on English, and students will quickly become familiar with the four corpora used most often in the studies: the London-Lund Corpus (LLC), the Lancaster-Oslo-Bergen (LOB) Corpus, the conversation register of the British National Corpus (BNC), and selected registers of the Longman-Lancaster Corpus.[2] The authors are explicit about the limits of the textbook: it is not designed to teach students computer programming, statistics, or corpus building. It does not provide a review of other scholarly approaches to contextualize corpus linguistics within the discipline; in fact, other than the lists of further reading, the text contains only a handful of references to outside scholarship. The focus of *CL* is how modern corpus linguistics, as a methodology, can aid language research; its aim is to provide a 'dynamic' perspective of this approach by describing specific, accessible sample analyses throughout the text.

This brief discussion of the book's potential audience highlights some pragmatic difficulties associated with the integration of an emerging field into the pedagogical realm. Many institutions, particularly in the US, have yet to devote a full course to corpus linguistics. As a result, it is unclear exactly where such a course would fit into the curriculum and, therefore, how much background the students would possess. Should corpus linguistics serve as a highly interactive, 'hands-on' approach to introducing lower-level students to the study of language? Or is it a more advanced methodological approach that students should acquire once they are familiar with the structure and use of language? BCR have written this textbook for the widest audience, with the 'lowest common denominator' linguistic background, which opens the possibility of teaching the text both within a linguistics curriculum and within an English curriculum. (For more advanced linguistics students, McEnery and Wilson (1996) or Stubbs (1996) may be more appropriate.) *CL* provides a clear introduction to how corpora can aid in language study, and throughout it raises intriguing questions for students to pursue. At the end of every chapter, the authors provide a brief list of further reading, which students should be encouraged to pursue to gain a wider understanding of the field. Instructors who employ this book as a base text will find it easy for students then to tackle specific studies within a particular area (eg syntax, register characterization), because they will understand the underlying premises and methods. Students should also be able to move comfortably from this introduction to more advanced study of, for example, programming, corpus design, or statistics.

The introductory chapter of *CL* frames the field, the tools, and the kinds of investigations possible. The body of the book is divided into two parts: the four chapters in Part I focus on investigating the use of particular language features; the three chapters in Part II focus on investigating the characteristics of varieties (eg registers, historical periods). The serious attention given in Part II to the study of varieties reflects the authors' impressive research in this area and distinguishes this book as an introductory text. Most treatments of corpus linguistics, and much of the research being pursued for that matter, focus on the association patterns of linguistic features; this book nicely highlights another set of possibilities for investigation. In both parts, the chapters are structured similarly: each begins by outlining the types of possible research questions in this area; the following subsections treat each type of question in detail by exploring an example analysis (all from English)—both the methodological steps and the interpretation of the findings. The concluding chapter briefly addresses teaching applications of corpus linguistic research. At the end of the text, the ten 'Methodology Boxes' provide short summaries of issues such as corpus design, tagging, and statistical methods, each followed by suggestions for further reading.

The Appendix lists available corpora, analytical tools, and various on-line resources.[3]

Chapters 2–4 build logically on each other, demonstrating the progression from simpler frequency and distribution searches to more complex analyses of, for example, verbal valencies. Chapter 2, 'Lexicography', leads students through the mechanics of frequency counts (including normed frequencies) to demonstrating how studies of written corpora can reveal important distinctions in the meaning and use of words. Specifically, the studies of *deal*, as well as *big*, *large*, and *great*, clearly show how collocates can distinguish the meaning of words, and how patterns of use can help differentiate 'near-synonyms'. This chapter devotes less attention to applications of these findings; for example, the discrepancy between these corpus results and dictionary entries is mentioned – as well as the importance of these findings for EFL teaching – but the conclusions about possible applications remain general: 'Lexicographic work should incorporate both perspectives, being complete in identifying the range of meanings but useful in marking senses that are most common or important' (p 41). Students should be intrigued by these issues, providing instructors the perfect opportunity to supplement the discussion with further reading and exercises designed to study applications of corpus research in lexicography.

Chapter 3, 'Grammar', focuses on how corpus linguistics can reveal the patterned ways in which speakers use the grammatical resources of a language; all the sample analyses in the chapter reinforce the importance of considering register in distribution patterns and the possibility of functional explanations based in part on communicative goals. The discussion of the distribution patterns of subject position and extraposed *that*-clauses is particularly helpful, with an interesting application to EFL teaching—the need to revise textbooks to reveal the rarity of subject *that*-clauses and to explain the reasons behind their use. Chapter 4, 'Lexico-grammar', synthesizes the previous two chapters in its examination of association patterns between words and syntactic constructions. The included analyses describe the different distribution patterns of the adjectives *little* and *small*, the different clausal structures following *start* and *begin*, and the different lexical associations with *that*- and *to*-clauses. Importantly, throughout these three chapters, the preliminary discussions of the studies highlight some of the methodological difficulties, including the 'fuzziness' involved in categorizing all the different types of complement clauses, in determining whether every adjective is predicative or attributive, and in identifying all nouns and verbs in any given text. It is crucial for students to recognize the human tagging process that underlies the 'hard numbers' in the final results and to learn about processes designed to enhance consistency (eg KWIC files).

Chapter 5, 'The Study of Discourse Characteristics', takes up the challenge of proving that corpus linguistics can be a productive approach to discourse analysis. Two particular strengths of this chapter are the examples of innovative output formats (eg the map of the progression of verbs through a text on p 128), and the explanation of an interactive identification program (pp 112–116). The two main example studies, while clearly explained, in many ways provide evidence for the intuitive, particularly about reference types in spoken and written discourse. The second study examines verb tense and voice patterns in the different sections of research articles; the systematic differences the study reveals could be used productively in teaching students to write more effective research articles—an idea the authors mention but do not actively pursue.

Chapter 6, 'Register Variation and English for Special Purposes', is a highlight of the book. After a highly accessible introduction to register, a short study reinforces for students the importance of questioning encompassing categories such as 'dependent clause' in any analysis. The authors then provide an impressively clear and concise explanation of multi-dimensional analysis, supplemented by examples of variation across spoken and written registers in general (eg conversation, fiction, academic prose) as well as within one particular register (eg different types of research articles). They plainly prove the effectiveness of multi-dimensional analysis, with its corpus-based approach, as a tool for providing more comprehensive characterizations of registers and for studying register variation; BCR also model how a final analysis should strive to identify the functional differences underlying the quantitative differences among registers.

Chapters 7 and 8, 'Language Acquisition and Development' and 'Historical and Stylistic Investigations', provide brief introductions to these areas, raising important questions that instructors should pursue with further reading and corpus studies. The studies in Chapter 7, based on the CHILDES database as well as two specialized corpora of elementary student language developed by Grabe and Reppen, track changes in students' speech and writing from third to sixth grade. Here, corpus-based approaches to studying language development clearly address some of the limitations of traditional case studies, enabling generalizations about, for example, the growing length of students' written sentences without a corresponding growth in clausal complexity, about the 'oral features' of student writing, and about students' control of patterns of register variation as early as the third grade. The final study examines which types of errors in native English-speaking and Navajo-speaking students' writing improve (and which do not) between third and sixth grade, suggesting the importance of corpus-based research in improving educational practices. It is only in a footnote, however, that the authors recognize that some of these 'errors' may be characteristics of

Navajo English and not 'errors' at all; given the controversy over how to negotiate Black English in the classroom and the known importance of recognizing dialectal difference in the pedagogical arena, this statement merits more than a footnote. Chapter 8 introduces the study of literary style and historical linguistics, with the recognition that the latter potentially includes all types of investigations discussed in earlier chapters. The sample diachronic studies provide highly preliminary investigations of the evolution of semi-modals, of medical research articles and drama, and of stylistic features in men's and women's letters. These studies are greatly simplified by the fact that they examine patterns of use only over the past three centuries; the authors, therefore, are not forced to address questions of spelling variation in earlier periods, let alone unfamiliar lexical and syntactic patterns. The authors rightly note that these historical results are not comprehensive, and it would be important for any instructor to supplement this section with more detailed historical studies, particularly ones that examine the intersection of extralinguistic factors (eg gender, class, educational background, age) in the development of linguistic change (cf Nevalainen and Raumolin-Brunberg 1996).

CL effectively introduces to students the types of questions corpus-based studies can address, reinforcing the potential gains offered by corpus linguistic methodologies coupled with rigorous critical analysis. The authors cover a wide range of language issues and types of corpus-based studies. The book in many ways raises as many new questions for students as it answers—which is perhaps a mark of a good introductory textbook. In the presentation of some examples, because the authors work to keep them basic and clear, the explanations of the methodology and of the statistical findings take away from attention to their implications. But instructors will have no problem rectifying this by challenging students to pursue possible applications of these results, by requiring students to read from the suggested reading lists, and by asking students to design their own follow-up studies. Importantly, *CL* provides students with clear examples of how such studies should be carried out and how the results should be analyzed. One stated goal of *CL* is to inspire students to pursue the questions the authors raise, as well as their own research questions; and the book explains and models for them a dynamic, accessible, and productive methodology for the pursuit of such linguistic knowledge.

Notes

1. The limited attention given directly to rationalist linguistics is perfectly exemplified by the fact that Chomsky's name appears nowhere in the text or bibliography of the book.

2. The authors' decision not to address dialectal differences between British and American English, while it simplifies the presentation of the corpus-based results, seems potentially problematic. As one example, these dialectal differences present an ongoing problem in EFL, an area the authors highlight as a site for the application of corpus-based study results.

3. The text and methodology boxes provide a good introduction to KWIC files, but it must be noted that nowhere in the discussion of tagging in *CL* is there mention of SGML or the new TEI guidelines.

References

Biber, Douglas and Edward Finegan. 1991. On the exploitation of computerized corpora in variation studies. In Karin Aijmer and Bengt Altenberg (eds) *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, 204–220. London, NY: Longman.

Fillmore, Charles J. 1992. 'Corpus linguistics' or 'Computer-aided armchair linguistics'. In Jan Svartvik (ed) *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82*, 35–60. Berlin, NY: Mouton.

McEnery, Tony and Andrew Wilson. 1996. *Corpus Linguistics*. Edinburgh: Edinburgh UP.

Nevalainen, Terttu and Helena Raumolin-Brunberg (eds). 1996. *Sociolinguistics and Language History: Studies Based on the Corpus of Early English Correspondence*. Amsterdam: Rodopi.

Stubbs, Michael. 1996. *Text and Corpus Analysis*. Oxford: Blackwell.

Svartvik, Jan. 1992. Corpus linguistics comes of age. In Jan Svartvik (ed) *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82*, 7–13. Berlin, NY: Mouton.

Sylviane Granger (ed). *Learner English on the Computer.* London: Longman, 1998. xxii + 228 pp. ISBN 0-582-29883-0. Reviewed by **Hilde Hasselgård**, University of Oslo.

Research on learner English is a comparatively new development within corpus linguistics. *Learner English on the Computer* offers a comprehensive survey of the field, bringing together research on various learner corpora from a range of nationalities and language backgrounds. Geoffrey Leech writes in his preface (p xvi) that 'The concept of a learner corpus is an idea whose hour has come'. After reading the book my impression is that the contribution of learner corpora is not only a timely one, but one which will soon prove invaluable in appreciating the structures of interlanguage and the process of foreign language learning, and in developing improved pedagogical tools and methods. Indeed, the book demonstrates that learner corpora are already at a stage where they incite research activity which in turn has practical applications.

Though several learner corpora are represented in *Learner English on the Computer*, the ICLE (International Corpus of Learner English) project is clearly the backdrop of the book. Sylviane Granger is not only the editor of the volume; she is also the (co-)author of four of the papers and the driving force behind the ICLE project. It is no doubt largely her merit that learner corpus research has gained the spread and the accomplishment that the book bears witness of. At present, ICLE comprises subcorpora from fourteen different language backgrounds. Each subcorpus, of approximately 200,000 words, consists of essays written by advanced students of English. In addition there is a 'reference' corpus of essays by students who are native speakers of English (the LOCNESS corpus, ie the Louvain Corpus of Native English Essays), which provides good opportunity for comparing the learner data with similar texts by British and American students.

This volume contains fifteen papers, plus a preface by Geoffrey Leech and an introduction by the editor. The papers have been grouped into three sections which are concerned with 'Learner corpus design and analysis', 'Studies of learner grammar, lexis and discourse', and 'Pedagogical applications of learner corpora', respectively. Together, parts I–III give a comprehensive and many-faceted picture of both second-language learning in general and the study and application of learner corpora in particular. Part III is followed by a list of the linguistic software described in the book, a bibliography and an index.

Part I is introduced by Sylviane Granger reflecting on the contribution of learner corpora in relation both to 'traditional' corpus linguistics and to Second Language Acquisition (SLA) research. Up to now, SLA research has mainly been based on introspective and elicited data, thus often failing to anticipate some learner problems. Research on authentic learner texts has sometimes yielded indeterminate results, because the material has not been controlled for factors such as register and level of proficiency. The use of native language corpora for English Language Teaching (ELT) purposes, on the other hand, cannot give any indication of the learners' difficulties with grammar or vocabulary. Granger then outlines the design of the ICLE subcorpora. The *shared features* of the subcorpora are age, learning context (university), level (advanced), medium (written), genre (essay), and technicality. *Variable features* are sex, mother tongue (L1), region, other foreign languages, practical experience, topic, and task setting. Evidently, it is vital to have a set of shared features in order to make comparison across L1 groups meaningful. Such comparisons will, for example, enable analysts to distinguish between general and L1-specific learning problems. Above all, the learner corpora, with the appropriate mark-up and software, will provide answers to research questions which could not previously be dealt with adequately.

In the next paper, Fanny Meunier describes some computer tools for the annotation and analysis of corpora. Since the ICLE corpus consists of advanced learners' English, part-of-speech taggers developed for native English perform well. With lower levels of proficiency, however, automatic tagging can be problematic. Of special value for learner corpora is the possibility of 'error tagging'. Such tagging must, however, be done manually for most types of errors, and is thus a time-consuming process. Areas for automatic analysis are suggested, such as sentence length, various features of lexis (eg frequency and variation), and some features of grammar and discourse.

Part II contains eight case studies. Though the foci of the papers vary, some points are common to several studies. One common pattern is learners' tendency to overuse or underuse lexical and grammatical structures, as compared to native speakers. Style is another recurring issue, as some of the patterns of over- and underuse suggest a different formality level from what is expected in an academic essay. The studies are predominantly quantitative, though interpretations of the patterns are often offered.

Håkan Ringbom examines the use of frequent vocabulary in corpora from seven L1 backgrounds. It is shown that non-native speakers (NNS) overuse some items and underuse others, as measured against frequencies in native speaker (NS) material. These discrepancies cannot be fully explained by reference to the learners' first language. Gunter Lorenz investigates adverbial intensification of adjectives by German learners and by English native speakers. Two sets of corpora are used, reflecting different age groups. The German learners in both groups are found not only to overuse adverbial intensification but also to use intensification in different contexts and for slightly different purposes than native speakers.

The study by Sylvie De Cock et al is the only one based on spoken English, by L1 French students and by English native speakers. The authors want to test whether learners use prefabricated sections of language to a lesser extent than native speakers. Their finding is, however that the NNS output contains nearly as many recurrent word combinations ('prefabs') as the NS material, but that the prefabs in the two corpora are not necessarily the same. They are also used for partly different purposes.

Bengt Altenberg and Marie Tapper examine adverbial connectors in the Swedish ICLE subcorpus as compared with the NS corpus. A corpus of Swedish L1 essays is also studied in order to control for L1 transfer. Transfer cannot, however, explain the divergence between the NS and the NNS materials. Instead it is suggested that the learners lack an awareness of formality and register, as they underuse formal features and overuse features typical of informal English. A tendency towards a too informal register is also found in Tuija Virtanen's investigation of the use of direct questions in essays. She finds that questions are used more often in NNS than in NS essays, thus reducing the argumentative effect. It is suggested that cultural differences in argumentative style may contribute to the differences. Virtanen also discusses the extent to which discourse studies can be done by computer. Computer searches are found useful for 'testing hunches', while the qualitative analysis requires interpretative work which cannot be done automatically.

Another discourse feature is explored by Stephanie Petch-Tyson, namely the visibility of the writer and the reader in the text. It is found that NNS essays (from five national backgrounds) contain more interactive features than the NS essays. The more 'involved' style of the NNS essays makes them seem more informal, though this is not commented on beyond the introduction. Formality level is, however, prominent in Sylviane Granger's and Paul Rayson's study of the automatic profiles of learner texts. Using tagged and lemmatized corpora (NS and L1 French NNS), they find patterns of over- and underuse of lexical items and grammatical forms in the NNS material that all point towards a less formal style than the one found in the NS material. Various explanations are suggested, eg the emphasis on speech in communicative language teaching.

Jan Aarts and Sylviane Granger use a tagged corpus (with the TOSCA-ICE tagset) to uncover differences in tag sequences between native and non-native speakers from three language backgrounds. An important aim of the study is to test the methodology. A list of the ten most frequent trigrams (sequences of three tags) in the NS material reveals similar patterns of over- and underuse in the NNS subcorpora, notably the underuse of patterns involving prepositions. However, the most frequent trigrams in the subcorpora show L1-specific features.

Part III shows how research on learner corpora may enhance ELT grammars, learner dictionaries, manuals for writing, and software for computer-assisted language learning. Doug Biber and Randi Reppen examine the use of complement clauses in the Longman Learner Corpus and the Longman Grammar Corpus. Traditional grammars usually give descriptions of the four types of complement clause, but little advice on how they are used. The key to that can be found partly in lexical association patterns and partly in register differences. It is thus a problem that the learner corpus is heterogeneous as regards register.

Patrick Gillard and Adam Gadsby report on the use of learner material in the compilation of the *Longman Essential Activator* (*LEA*). They claim that, just as the use of native language corpora has greatly improved learners' dictionaries, the use of learner material will enhance the value of dictionaries chiefly aimed at intermediate learners. Such dictionaries can take into account the kind of vocabulary the students know and the areas in which they are likely to go wrong. One of the features of *LEA* is thus 'help boxes', which warn learners explicitly against common errors.

The concern of Przemyslaw Kaszubski's paper is English writing textbooks. Existing material is often aimed at native speakers, and does not cater for the special needs of learners. Intuition-based textbooks for learners may likewise fail to address the appropriate problem areas of lexis and grammar. Corpus-based L1-specific writing manuals would be able to focus on errors that commonly occur in the writing of (in this case) Polish learners. Likewise, the comparison of NNS writing with NS writing will uncover areas of lexical and grammatical overuse and underuse. Learners could then be advised accordingly.

John Milton describes a software package for Cantonese-speaking learners of English, based on evidence from a corpus of learner English from Hong Kong. Preliminary research included error tagging of the learner corpus in order to uncover the most common error types. Word sequences were compared to a NS corpus to survey areas of over- and underuse. The software package contains an on-line grammar and an editing/proofreading tool which picks up on the common types of errors and has direct access to the grammar. Further, it has a list-driven concordancer of underused word sequences to increase the learners' repertoire of expressions, since a common problem among these learners is massive overuse of a limited number of expressions – a feature which is clearly teaching-induced.

Finally, Sylviane Granger and Chris Tribble discuss the direct use of learner corpora in the classroom. They start from an increasing awareness of the value of corrective feedback and the importance of accuracy within communicative language teaching. It is suggested that NNS material can be used, albeit with caution, in exploratory exercises, using short concordances which show for instance, the complementation of verbs in NS and NNS material, or alternatives to vocabulary items which learners overuse. The question of what type of NS material is best suited for reference for the learners is also given some attention.

The papers included in *Learner English on Computer* show many intriguing aspects of learner corpora, and the book can be read either as an introduction to learner corpus design and research or as an argument for doing such research. However, the large number of papers is also problematic, in that it severely limits the length of each study, and thus the possibility of the contributors to discuss their findings in some more detail, rather than scratching the surface.

A question that remains unanswered is whether corpus linguistics and SLA have really met in learner corpus research. While learner language corpus research does not seem to be very controversial in relation to traditional corpus linguistics, some potential conflicts in relation to traditional SLA/ELT research are hinted at, but these conflicts are not resolved, nor commented on by anyone from 'the other side'. As it may be unfair to criticize such a rich volume for what it does not contain, we may instead anticipate a sequel, presenting further applications of learner corpora with results of new methods and improved tools in foreign language teaching.

Learner English on Computer has an internal coherence, which makes it more than a collection of papers. It is carefully planned, with a structure that takes the reader from the background for learner corpora via corpus design to (preliminary) research results and applications of these. The book is thus a landmark in asserting learner corpus linguistics as a field of research in its own right, and is recommended to anyone with a (potential) interest in learner language.

Stig Johansson and **Signe Oksefjell** (eds). *Corpora and Cross-linguistic Research. Theory, Method and Case Studies.* Amsterdam and Atlanta, GA: Rodopi, 1998. 376 pp. ISBN 90-420-0291-3. Reviewed by **Fanny Meunier**, Université Catholique de Louvain, Belgium.

Corpora and Cross-linguistic Research is an edited volume of research articles written within the framework of the 'Contrastive Analysis and Translation Studies Linked to Text Corpora', a project directed by Stig Johansson at the Centre for Advanced Study, the Norwegian Academy of Science and Letters, Oslo. And in fact, although it is not a proceedings volume, all the contributors to the book attended the Fourth Nordic Symposium on Text-based Contrastive Studies in Oslo in 1997. It is a substantial book, divided into two main sections: the first, smaller, contains five articles on theory and method and the second, larger, comprises ten case studies. The collection is a perfect example of 'second generation' cross-linguistic studies: relying on past experience, taking advantage of new parallel computer corpora and technologies, providing results based on firm evidence from corpora and drawing conclusions for both applied and theoretical research.

After Jan Aarts' introduction, Stig Johansson's opening article discusses the relative advantages and disadvantages of five types of corpora used in cross-linguistic research. It also provides a comprehensive introduction to the English-Norwegian Parallel Corpus (or ENPC, which has provided the impetus for much of the research in this book), describing the structure of the corpus, the research methodology and navigation procedures, which are then further illustrated with the help of three concrete examples. The examples show how corpora may serve 'to unite fields of study that have traditionally been kept apart' (p 21) and result in better, more coherent language description. Martha Thunes' article addresses the notion of complexity in translation. She distinguishes four main types of translational correspondences of increasing complexity, using a 'corpus' of finite clauses found in the ENPC. Then Helge Dyvik stresses the important role of translation corpora in the development of linguistic semantics, which 'traditionally is heavily monolingual in scope' (p 51). He examines the translational properties of three Norwegian lexemes in the ENPC and demonstrates how the results of translational studies 'should be seen as imposing constraints on the set of possible semantic models for the languages'.

The last two articles in the first section review software programs, both developed within the framework of the ENPC project. Knut Hofland and Stig Johansson present the Translation Corpus Aligner, a program for the automatic alignment of parallel texts. Unlike similar, chiefly statistics-based programs, the particular interest of the Translation Corpus Aligner (TCA) lies in its focus on language specific information in the form of 'anchor words' (the selection of which was based partly on intuition and partly on manual matching of original and translated texts). The TCA is also able to automatically extract cognates, based on the principles used in the CRATER project (McEnery and Oakes 1995). It has been tested on a total of 93,000 sentences (1.3 million words) with an average error rate of 1.98 per cent, making it a very accurate and robust pro-

gram which may also, in the future, be further developed to carry out automatic word alignment, as Hofland and Johansson discuss. The program presented by Jarle Ebeling in the final article is the Translation Corpus Explorer (TCE), a browser for parallel texts. All ENPC texts have a file header information file with structural (paragraphs, sentences, etc) and descriptive (title, author, text type, etc) mark-up. The TCE operates by means of a database generated from the text and is able to carry out lexical word-based searches (single and multi-word) and use wildcards and Boolean operators. The position of the words in the sentence can also be used as a search option. The TCE was originally designed to handle pairs of texts but it is now possible to view the same original texts, translated into several languages. Finally, Ebeling discusses future extensions to the program, such as part-of-speech tagging and extended filtering possibilities.
The ten articles in section two of the volume cover a range of lexical, grammatical and discourse topics. Bengt Altenberg's study of connectors and sentence openings in English and Swedish demonstrates how a contrastive approach to grammatical differences can reveal interesting discourse features and writing strategies. Hilde Hasselgård studies English/Norwegian translation pairs in which word order changes result in a change of thematic perspectives, or where thematic structure is retained despite syntactic restructuring in the translation. She argues that if a slight alteration is made in the text structure, the resulting translation may make a slightly different impact on the reader, even if there is no alteration in the themes and referential meaning.

In his second contribution to the volume, Jarle Ebeling studies the English existential *there*-construction, analysing Norwegian target language expressions that capture as many of the features of the source construction as possible. If no similar structure is used in the translation, he then tries to define what features of the original are deemed most salient. Most of the time the existential *there*-construction is translated by its Norwegian counterpart, the *det*-construction. When the *det*-construction is not used, the discourse function of the English *there*-construction seems to prevail, followed by the impersonal feature, with the aspect of asserting existence as least important.

Cathrine Fabricius-Hansen's concern is information splitting and its effect on discourse structure. Using a corpus of German non-fictional prose (which presents high informational density and a high degree of syntactic complexity) translated into English and Norwegian, she introduces the notions of hierarchical vs incremental discourse information structure and shows that, although the English and Norwegian translations both increase incrementally, they do so in different ways. Whereas the English translation compensates for the information splitting with a refined register of clause combining devices at sentence and discourse level, the Norwegian translation uses extensive information splitting without overt compensation at a higher level of sentence or discourse structure. The topic of Monika Doherty's paper is the empirically identifiable constraints on phrasal reductions in German translations of English adverbial clauses. Her corpus investigation of *when*-clauses reveals that more than fifty per cent of the English clauses in medial or final positions were translated as phrases. She demonstrates that such reductions may be both syntactically and semantically, even sometimes stylistically, motivated. The following article by Josef Schmied also deals with English and German, examining translation correspondences between the close cognates *with* and *mit*. In fact, few are revealed, which leads him to draw conclusions for lexicography, translation aids, language teaching and language change.

The next two papers in the section both investigate the phrase *I think*. Karin Aijmer deals primarily with its Swedish and German translation counterparts and Anne-Marie Simon Vandenbergen with its translation into Dutch. Karin Aijmer illustrates the polysemy of *I think* and discusses three strategies for expressing subjective evaluation. She also demonstrates the contribution of cross-linguistic studies to a deeper understanding of the system of present-day English. Anne-Marie Simon Vandenbergen investigates the semantic and pragmatic functions of *I think* in English and Dutch spontaneous conversations and parliamentary debates. She comments on the English *I think*, on the comparison between English and Dutch and on the factors influencing the choice of translation equivalents.

In the next article, a study of perception verbs in English and Portuguese, Diana Santos rejects contrastive approaches which start from an assumption of a priori universal or common features. The role of contrastive studies, she says, is 'to find similarities or differences, not to presume them from the start' (p 319). She first compares the properties of perception verbs in English and Portuguese and then discusses their respective translations, highlighting the frequent addition or omission of perception verbs in translations. She also shows how her data can be reinterpreted in the light of previous research results. In the final paper of the volume, Åke Viberg studies the highly polysemous English verbs *run* and *put* and their Swedish counterparts. He makes some interesting comments on the notions of differentiation and neutralization in translation and examines the syntactic, semantic and pragmatic clues crucial to the interpretation of these verbs.

The fifteen articles contained in *Corpora and Cross-linguistic Research* successfully demonstrate the wide variety of linguistic fields in which parallel corpora are currently being used, and highlight the vast potential of modern cross-linguistic research, particularly by making explicit the link between practical investigation and theoretical considerations, in keeping with a belief fast gaining ground in corpus linguistics that corpus work has an important part to play in 'theoretical linguistics'. The book offers a highly comprehensive selection of writing from top scholars on theory, methodology and quantitative and qualitative approaches in cross-linguistic research and, as such, is a must for anyone interested in the field.

Magnus Ljung (ed). *Corpus-based Studies in English. Papers from the Seventeenth International Conference on English Language Research on Computerized Corpora (ICAME 17), Stockholm, May 15–19, 1996*. Amsterdam, Atlanta, Ga: Rodopi, 1997. 388 pp. ISBN 90-420-0341-3. Reviewed by **Gerald Nelson**, Survey of English Usage, University College London.

This volume is one of the largest collections of ICAME conference proceedings, with a total of twenty-four papers on a very wide range of topics. The editor has used the three-part format which seems to have become established practice in publishing these proceedings: Part 1 contains just two papers, under the general title Parallel corpora and translation studies. The main part of the book is Part 2, Synchronic and diachronic studies of English, which contains eighteen papers. These studies present the results of a wide range of corpus-based research, as well as some more theoretical papers on methodology. The four papers in Part 3, Parsing and tagging, look at specific software solutions to problems in grammatical analysis and lemmatization.

Many of the most interesting papers address issues in the methodology of corpus linguistics, and might usefully have been collected in a separate section. William Kretzschmar, Charles Meyer, and Dominique Ingegneri wonder whether it can ever be logistically possible to compile a truly representative corpus of American English, given the ethnic, geographical, and linguistic diversity of the US. They begin by looking at sampling methods used in other fields, specifically by social scientists conducting political surveys. Applying one of these sampling formulae, the authors calculate how many books published in the US in a given year would have to be sampled in order to ensure a statistically representative corpus of American writing. The results are not encouraging. They calculate that not less than two thousand books would have to be sampled. Even if only 2,000 words are selected from each book, this would mean a corpus of over 4 million words, far bigger, as the authors point out, than either Brown or LOB.

They conclude that 'logistically' the best a linguist can hope to achieve is a corpus which is 'reflective', rather than 'representative', of a variety of English (p 168).

In the final part of this paper, Kretzschmar et al discuss in broad terms the use of significance tests in analysing corpus data. Unfortunately, they begin their discussion as follows: 'After having gathered an acceptable sample of a population...' (p 173). I have no idea what an 'acceptable' sample is, or why the authors have suddenly abandoned the terms 'representative' and 'reflective'. In any case, the subsequent discussion does not offer any new insights, and simply reiterates common knowledge in corpus linguistics: our results are contingent on the corpus we use, and we should select significance tests which are appropriate to the type of data we are examining.

The use of statistics is discussed more cogently by Christian Mair, in his paper on grammatical change. Mair writes: '....the relationship between statistical and linguistic significance is a complex one: there are statistically significant patterns in corpora which cannot be interpreted linguistically, and some linguistically significant facts from corpora are not statistical' (p 201).

This is written in the context of a paper which attempts to find a middle ground between corpus linguistics, in which large bodies of evidence are available, and grammaticalization theory, which relies on a small number of examples, and lacks evidence once the focus switches to recent phenomena. Mair believes that both of these disciplines have something to offer the other. For corpus linguistics, grammaticalization theory offers the opportunity to consider a wider view, which may be supported by, but not bound to, corpus evidence and statistical significance. Mair writes candidly: 'Corpus-linguists sometimes go about their business with a degree of naiveness, happily adding a further set of statistics to existing counts, and not worrying too much about the broader significance of such findings' (p 198). This is a wide-ranging paper, which also reports the results of a study of the increased 'informality' of written English since the 1960s. Mair argues that increasing use of the progressive, of the *going-to* future, and of contracted forms in the last thirty years cannot be interpreted as changes in the grammar. Rather, the evidence suggests that these informal options, which have been available for a long time, are chosen more frequently today than they were in the past. Mair goes on to consider these developments as the linguistic correlates of a more general informalisation of codes and manners in society in the same period.

Inge de Mönnink is also concerned with methodology, and with the limitations of corpus data. She focuses on certain types of 'non-regular' NPs (eg *we both*, *so romantic a name*, p 227), which generally yield very few instances in a corpus, but are typically produced by native speakers. She calls for a judicious use of elicitation tests in these cases, as a means of supplementing corpus data. Her paper is a detailed examination of how elicitation tests should be designed and implemented. She concludes that it is very difficult to design good elicitation tests, ones which properly control for all possible external variables. Nonetheless, the results of these tests, together with native-speaker acceptability judgements, have an important contribution to make to the description of language use.

This volume is notable for the number of different corpora which have been examined by contributors. As well as the more ‘traditional’ datasets, such as Brown, LOB, and Helsinki, several recently compiled corpora are also represented, including COLT (the Bergen Corpus of London Teenage Language), ICE (the International Corpus of English), and the Lampeter Corpus of English pamphlets, 1640–1740.

The COLT corpus has yielded many valuable publications, two of which are included in this volume. Gisle Andersen’s contribution looks at the sociolinguistics, grammar, and pragmatics of the discourse marker *like*, while Hasund and Stenström examine the sociolinguistics of verbal disputes among teenage girls. One of the most valuable features of this corpus is the coding of informants for social class; both of these papers exploit this information.

Vincent Ooi examines the Singapore component of ICE for lexicographic evidence of English in a second-language context, and considers arguments for the inclusion of ‘regionalisms’ in learners’ dictionaries. Despite its small size (one million words), ICE-Singapore proves to be a useful starting-point for this kind of study. Ooi offers a conceptual framework for controlling the inclusion of items in a learners’ dictionary, based on five types of words – from ‘core’ English to informal, usually spoken, ‘Singlish’.

A number of papers report the results of diachronic studies. Using part of the Lampeter Corpus, Claudia Claridge compares the use of multi-word verbs in two decades, the 1640s and the 1730s. Thomas Kohnen uses data from the Helsinki Corpus to consider the proposition that the evolution of a language may be conceived as a history of its text types. Focusing first on historical writing, Kohnen considers the development of this genre in terms of the use of *-ing* participle constructions. Among other findings, Kohnen reports a sharp increase in the use of these constructions in the 15th century, followed by a sharp decrease from 1650 onwards. He goes on to look at the same construction in other text types, including administrative prose, fiction, and private letters. Perhaps not surprisingly, he finds a similar overall pattern for all four types, though the increase is less marked in personal letters. However, the most interesting result is that each genre follows its own chronological pattern. The increase in participle constructions occurs earliest in administrative prose, followed by historical writing, fiction, and letters, in that chronological order. Kohnen concludes that, in quantitative terms at least, participle constructions develop according to text type, and more generally, that text type is a crucial factor in the spread of grammatical constructions. He is careful to point out that many more genres and textual functions would need to be studied to support this type of study, though on the evidence presented here, it is certainly worth exploring further.

The editor of this volume clearly had a difficult task. The growing popularity of the ICAME conferences, and the constant expansion of the field, have meant that he was presented with a very large number of submissions on a very wide range of topics. In the tradition of ICAME, he has chosen to be as inclusive as possible. However, this inclusiveness has meant that not enough space has been left for the graphics. In particular, Pieter de Haan's study of the syntactic characteristics of dialogue and non-dialogue is not well served by the quality of reproduction of his bar charts (pp 107ff). Similarly, Nancy Belmore presents her comparison of two taggers using pie charts (p 337), which are, unfortunately, far too small and indecipherable. No doubt all of the graphics in this volume began life as transparencies on an overhead projector at the Stockholm conference. I hope they were more legible there than they are in the printed version.

This volume provides ample evidence of the expansion of corpus linguistics in recent years. New corpora and software tools are becoming available all the time, and many new areas are being opened up to corpus-based study. At the same time, it is encouraging that this volume still finds room for more theoretical papers, such as those by Mair and by Kretzschmar et al, which force us to reflect on some of the basic assumptions and procedures of corpus linguistics.

Terttu Nevalainen and **Leena Kahlas-Tarkka** (eds). *To Explain the Present: Studies in the Changing English Language in Honour of Matti Rissanen*. Helsinki: Société Néophilologique, Helsinki (Mémoires de la Société Néophilologique de Helsinki, 52), 1997. xix + 503 pp. ISBN 951-96030-6-9. Reviewed by **David Denison**, University of Manchester.

The *tabula gratulatoria* reveals worldwide respect for the indefatigible birthday boy, Matti Rissanen — an implausible 60 year old at the time of publication — who has done so much over many years to stimulate and promote variation studies and diachronic studies in English corpus linguistics, and who has made the

Helsinki English Department the place to do it. (The recent designation by the Academy of Finland of the Research Unit on Variation and Change as a Centre of Excellence is a national recognition of his and his colleagues' achievement.)

This Festschrift contains a brief introduction, a list of Matti Rissanen's publications, and twenty-nine very varied papers by some of his friends and colleagues. There is no index. The editors had an invidious job selecting their thirty-five contributors (six papers are jointly authored), and the reviewer pressed for space is also forced to be selective in his comments, here largely descriptive. The collection is heterogeneous, and I cannot go much beyond a mere listing. They are divided by the editors into two rough-and-ready chronological groups, pre- and post-Early Modern English, and are arranged within each group alphabetically by first author. There are synchronic papers and diachronic papers, variationist papers, a lot of corpus linguistics, some more literary-philological contributions, a few theoretical or methodological disquisitions, and papers which partake of several of these headings, or none. So many strands are paid out and picked up at different points in the volume that one could weave one's way perfectly soberly through a review in all sorts of different orders. Here we go, then.

Matti Rissanen is of course closely associated with the Helsinki Corpus (henceforth HC), if not its *onlie begetter*, and historical corpus linguistics plays a large role in the collection. Only a few contributors make direct and exclusive use of HC, however. Matti Kilpiö discusses the history in English of participial adjectives like *said*, *aforementioned* used as anaphoric discourse and/or style markers. Kirsti Peitsara starts from HC but brings in much other evidence in her detailed study of the syntax of *enough* in Middle English and of the origins, distribution and significance of variation between the form types ENOW and ENOUGH. Antoinette Renouf experiments on the HC with software developed to detect lexical innovations by means of collocating words. Although the software was designed for a corpus of current English, to which newer and newer blocks of text are continually being added, she sees some possibilities in her case studies for detecting semantic or grammatical change in a historical corpus.

The Helsinki Corpus has spawned a number of other collections of historical English material. From the data in her Helsinki Corpus of Older Scots, Anneli Meurman-Solin provides a substantial examination of *t/d*-deletion. Irma Taavitsainen and Päivi Pahta introduce another corpus conceived out of HC, The Corpus of Early English Medical Writing, with a preliminary study of phrases of the type *It is to V.* Helena Raumolin-Brunberg and Arja Nurmi use yet another, The Corpus of Early English Correspondence, to focus on two dummy elements, nominal *one* and verbal DO. Their interests are register variation and the social status of individuals leading a change.

Merja Kytö and Suzanne Romaine use the BNC plus ARCHER and HC to plot the rivalry between synthetic or inflectional (*quicker*) and analytic or periphrastic (*more quick*) comparatives in Modern English, likewise superlatives, plus the occasional use of synthetic and analytic types together in double comparatives. In this preliminary study there is an important discussion of the path of chronological change, illuminating key stages with a number of figures and charts, though the tentative conclusions on the significance of such factors as word length, word-ending and text type are not unexpected. Geoffrey Leech and Jonathan Culpeper tackle the same topic for recent British English using BNC and other sources. The chronological range is much narrower, but the analysis (of comparatives alone until section 3) is pursued a little more deeply, with syntactic function given greater attention. Manfred Görlach collects an impressive body of data from various sources on those strange *than whom/than which* comparatives, which may have been based originally on a Latin model and which later gave such trouble to prescriptivists.

We have now moved on to diachronic studies of texts and corpora other than HC. Douglas Biber and Edward Finegan correct their own earlier work by using ARCHER, their extensive historical ModE corpus, to show that 'specialist expository registers … have followed a consistent course towards ever more literate styles' (p 273) over the last 350 years, whereas it was only the popular written registers that showed a reversal towards more oral styles. Norman Blake promotes his *Canterbury Tales* project to test Fisher's history of Chancery Standard. Although the paper is concerned mostly with spelling variants, an excursion into the syntax of negation implies that *ne* as proclitic verb negator is being counted with the very different conjunction (p 16). Saara Nevanlinna examines lexical variation in OE Gospel manuscripts, concluding with some comments on the general linguistic situation at the time of late OE and early ME copying. Risto Hiltunen examines syntactic and textual construction in the Anglo-Saxon *Laws*. Bruce Mitchell has an intricate discussion of unexpressed principal clauses in Old English, referring in the main to previously-published but scattered comments. Michiko Ogura has an immensely detailed tabulation of *faran/ feran* variation in OE and early ME.

Coins are not a text-type represented in the Helsinki Corpus, and it would be stretching a point to call the decipherment and interpretation of one word — the name of an Anglo-Saxon coiner — corpus linguistics. Fran Colman's paper is closer to traditional philology, and she offers an etymological and phonological discussion of what is probably *Tilred*. Like Matti Rissanen, Antonette diPaolo

Healey has an interest in computational assistance for the mapping of the history of English. She applies it to the Toronto *Dictionary of Old English project*, discussing the background of the project and focusing on the semantics of *eald*. Moving more towards literary history, Fred C. Robinson re-examines the meaning of, among others, *fæhðe ond fyrena* in a close reading of *Beowulf* 879a. For other literary-based papers we jump forward again from Old English to the beginnings of the late Modern English period. Susan Wright [now Fitzmaurice] looks at relative markers in the writing of Joseph Addison, while Ingrid Tieken-Boon van Ostade examines the prescriptive work of Robert Lowth to find the texts which he had used to exemplify either good usage or the occasional lapse of a 'best' writer into error. She goes on to test Lowth's own letters against his later published precepts.

Three substantial papers tackle methodological or theoretical issues. Dieter Kastovsky examines the theoretical basis of morphological classification in (especially) Old English and warns against confusion between etymologically and synchronically justified labels. Roger Lass takes apart the paradigmatic evidence for the gender of OE *hus* 'house', cloaking his sophisticated theoretical musings in, characteristically enough, cheerfully relaxed language. He concludes that variation is ubiquitous and important, and that assignment of nouns to particular genders and declensions is often a matter of 'mostly' rather than 'definit(iv)ely'. John Anderson looks at the theoretical status of potential auxiliaries in English (mainly Present-day) as well as cross-linguistically, attempting at length to disentangle the concepts of morphosyntactic and syntactic auxiliary, where the latter is the narrower category. This is a demanding paper, presented as a preliminary to a study of auxiliarisation in English.

In another preliminary, if less ambitious, paper, Barbara Kryk-Kastovsky considers items like *now* used as discourse particles instead of temporally in English, German, Polish and other Slavic languages, arguing from introspection and from a brief inspection of historical dictionaries that there has been a cross-linguistic process of grammaticalisation. Mats Rydén searches for a core meaning to the English progressive, citing some diachronic observations but where possible using PDE to stand 'panchronically' for all periods.

Another strand of this volume runs through some synchronic studies of Present-Day English. Magnus Ljung tackles adverbial clauses headed by a subordinator and with a non-finite verb or no verb at all: *When entering his house, While in Paris*, etc. The corpus is a varied PDE one, and the investigation follows Biber in trying to find differences in usage according to genre. Stig Johansson uses four registers in (apparently) his own large corpus of PDE to look for discourse functions of existential *there* clauses. Gunnel Tottie examines choice of relative marker in the BNC, LLC and Corpus of Spoken American English (UCSB), plotting the distribution of markers in restrictive relatives against syntactic function and nature of antecedent. One important conclusion is that educated British English is not typical of contemporary English generally.

A paper which stands rather apart from the others
is the one by Jan Svartvik and Alex Chengyu Fang,
who experiment with the SpeechMaker software
to help non-native speakers
divide their written texts
into appropriate prosodic 'chunks'
for public delivery. [1]

In fact the research involves conventional but serious problems of tagging and parsing, here applied in an unexpected way.

Overall, then, readers of this Festschrift will find that it ranges from subtle theoretical analyses and comprehensive collections of interesting data to preliminary case studies, programmatic pieces and squibs. The general quality is high, the editing is careful, and the range of topics is wide. Many Happy Returns, Matti.

Notes

1. The lineation is by David Denison. – The Editors

Antoinette Renouf (ed). *Explorations in Corpus Linguistics*. (Language and Computers: Studies in Practical Linguistics, 23.) Amsterdam – Atlanta, GA: Rodopi, 1998. 292 pp. ISBN Hb: 90-420-0751-6, Pb: 90-420-0741-9. Reviewed by **Kay Wikberg**, University of Oslo.

The editor's excellent preface to this interesting selection of papers from the XVIIIth ICAME conference held in Chester in May 1997 would very well do as a review of the book. Antoinette Renouf groups the papers into three sections: 'Corpus creation: methods and issues' (two papers), 'Corpus analysis: synchronic and diachronic studies' (sixteen papers), and 'Corpus linguistic results: creation of resources and tools' (three papers). Thus, the bulk of the research represented in this volume is as usual concerned with 'corpus analysis', or to use a concept that may be more illuminating, 'corpus-based descriptions of English'. 'Analysis' primarily refers to processes, and the distinction Greame Kennedy (1998) makes between 'corpus-based descriptions of English', on the one hand, and 'corpus analysis', on the other hand, is therefore justified. Under 'corpus analysis', he deals with such processes and procedures as annotation and processing, listing, sorting, counting, and concordancing.

I will not necessarily review the papers in the order in which they have been included but find it appropriate to start with the first paper, 'Protecting the innocent: The issue of informants' anonymity in the COLT corpus' by Kristine Hasund. She is concerned with the legal and ethical aspects of compiling corpora of non-surreptitiously recorded speech but also with sociolinguistic and computational considerations. One major recommendation based on experiences with COLT is that first names should be preserved. Another conclusion she makes is that 'informants' rights to privacy should and must override other concerns' (p 25).

Gunnar Bergh, Aimo Seppänen and Joe Trotta contrast the 'standardised corpora' with the open-ended free text corpora available on the Internet. Using a rather rare syntactic construction (These are super-light particles *which* are believed can fill any space), they demonstrate the importance of corpus size for attesting grammaticality at the same time as they point at the lack of search tools for concordancing Internet material. The paper no doubt signals the interest of many researchers who would like to go beyond conventional corpora in their search for data. The Gothenburgh team's paper raises many questions, some of which I am sure will be answered in the next century.

The ICLE (International Corpus of Learner English) project, directed by Sylviane Granger, provides the basis for two papers, Pieter de Haan's 'How 'native-like' are advanced learners of English?' and Håkan Ringbom's 'High-frequency verbs in the ICLE corpus'. Pieter de Haan looks at word class sequences (eg Prep-Art-N, N-Prep-Art-N) in four subsets representing essays written by Dutch, Finnish, French and English (native speakers) students. To be able to interpret the results of this investigation, one would have liked to know more about the possible transfer effect of word order in the languages of the non-native speakers and also something about typical lexical realisations of the various word class sequences in the essays. Data from the ICLE corpus allow Ringbom to show how non-native speakers overuse high-frequency verbs like *think* and *get* at the expense of less common verbs, resulting in 'insufficient and imprecise use of the resources available in English' (p 199). Another finding is that 'the verbs most used by learners tend to be much the same regardless of the learners' L1' (p 194).

Using two dictionaries, one for Old and one for Modern English, Göran Kjellmer makes a diachronic study of initial bipartite consonant clusters. This is a valuable contribution to our understanding of English phonotactic structure. Kjellmer explains the existence of Present-day English consonant clusters as the result through time of 'an effort to maximise phonological contrast, and [...] a desire to simplify consonant sequences by cluster reduction' (p 94).

Several papers address syntactic problems. Christine Johansson and Christer Geisler investigate 'Pied piping in spoken English' (a fronted preposition + *wh*-word as in *The poem of which I spoke*), using the London-Lund Corpus, the spoken component of the Birmingham Corpus and parts of the BNC. They show that Prep+*which* is 'far more frequent in speech than is generally assumed' (p 79), and that it is preferred to the stranded preposition when the preposition + relativizer has an adverbial function. A related topic is discussed by Christian Mair in '*Man/woman which* ... – Last of the old, or first of the new?', which focuses on *which* with human antecedents in Early Modern English, the BNC, and data of his own. Mair concludes that it is not possible to detect any change in the use of *which*. The uses he has found seem to be archaisms or due to failure to define the feature [human] properly. Another study of English syntax is Magnus Levin's, 'On concord with collective nouns in English'. Levin uses spoken and written data from the BNC, in all some 54 million words. He finds that it is hard to generalize about concord, owing to the effect of a large number of variables, including specific nouns, text types, the presence of postmodifiers, and so forth. Nelleke Oostdijk assigns her study of the register of air travel information ('Language use in a restricted domain') to the field of computational linguistics rather than descriptive corpus linguistics. I find this classification unmotivated, since any study of register is also part of the description of a particular language, as Halliday and other functional grammarians have documented many times. Finally, Andrea Sand looks at 'First findings from ICE-Jamaica: The verb phrase'. Her comparative pilot study, based on a spoken and a written subcorpus of ICE-Jamaica and equivalent text categories from the BNC, FLOB and Frown, shows the importance of this kind of research. One specific finding is that backshifting is still the rule in British news data, whereas it is not a regular feature in Jamaican usage.

Apart from Ringbom's study of word frequencies in learner language, there are three other papers on lexical or lexico-grammatical topics. Using the tagged LOB corpus and the CD-ROM-version of the OED, Manfred Markus shows in '*A*-adjectives (*asleep* etc) in postnominal position: Etymology as a cause of word order (corpus-based)' that *a*-adjectives make up a mixed group etymologically, and that it is only the subset derived from *on/in/of* + Noun Phrase that have a significant number of postnominal occurrences. The editor herself and R. Harald Baayen continue their work on neologisms in 'Aviating among the hapax legomena: Morphological grammaticalisation in current British newspaper English', which focuses on adjectival derivations involving *-type, mock-* and *-shape* occurring in *The Independent* in the period 1988–1997. It turns out that most of these formations are ephemeral and typically meet the pragmatic or stylistic needs of journalism. Hence, both *-type* and *-shape* affixations are useful when precise information is missing or when there is no need for it. Finally, Pam Peters' 'In quest of international English: Mapping the levels of regional divergence' describes a relational database of parallel terms in Australian, British and American English. In her search for a method of pinning down international English, she gives full credit to the glossary by Hofland and Johansson (1982), which she claims is still very useful for her particular purpose, thanks to its comparative frequency lists.

The ACRONYM project carried out in the Research and Development Unit at the University of Liverpool has already been presented at several ICAME conferences. By plotting the collocational profiles of specific lexical items, it is possible to establish a set of 'nyms', ie words occurring in similar collocational environments. Alex Collier writes on 'Identifying diachronic change in semantic relations', where diachrony comprises semantic change in the last decade or so. The paper is interesting from a methodological point of view. A useful spinoff from the ACRONYM project is their new words service available on its web site. Mike Pacey's paper, 'The use of clustering techniques', deals with much more technical aspects of clustering.

Bas Aarts, Gerald Nelson and Justin Buckley report on 'The Internet Grammar of English: New horizons in grammar pedagogy', an online grammar for self study. This is a welcome resource at a time when more and more university institutions are planning to make their courses available on web sites to students anywhere and at any time. This Internet Grammar has been completed and can now be tried out on the Internet or bought in a CD-ROM version.

Three papers are based on the Corpus of Early English Correspondence, compiled at the University of Helsinki. Minna Nevala's '*By him that loves you*: Address forms in letters written to 16th-century aspirers' applies politeness theory to the analysis of address forms to three aspirers. The author shows that variation in the use of address forms can be interpreted as indicators of social movement. There is also evidence that 'the recipient's rise had its effect on the increase in the amount of negative politeness in the address forms' (p 157). The data used by Arja Nurmi in ‘Periphrastic DO and the language of social aspirers: Evidence from the Corpus of Early English Correspondence’ consist of nearly 2,300 instances of DO in the part of the corpus that represents the 16th century. She shows that periphrastic DO was more common in informal letters in the first half of the century, but that the construction gained ground in formal contexts in the latter half of the century. However, it did not become popular with social aspirers. Jukka Keränen’s ‘The Corpus of Early English Correspondence: Progress report’ is what it says, a progress report. The extralinguistic variables taken into account in the process of data selection include the writer’s provenance, social and family status, sex, education, age, and relation to the recipient. The size of this historical sociolinguistic corpus will be over 2,5 million words with nearly 800 informants. There are still copyright problems to sort out, but the project has already generated an impressive list of publications.

Finally, a note on two papers dealing with different aspects of discourse. Anne Wichman, in ‘Using intonation to create conversational space: Projecting topics and returns’, shows how a citation contour, consisting minimally of a falling tone, can occur at points which are not syntactically complete and have an organising role in discourse. Martin Wynne, Mick Short and Elena Semino report on ‘A corpus-based investigation of speech, thought and writing presentation (ST&WP) in English narrative texts’. What is new about their research is that they have manually tagged a corpus of some 250,000 words for types of ST&WP. In their paper, they describe problems that they had with the tagging, such as dealing with ambiguity. A result of their detailed analysis is that they also had to add a few new categories such as ‘narrator's report of voice’ and ‘narration of internal states’ to the set originally described in Leech and Short (1981).

References

Hofland, Knut and Stig Johansson. 1982. *Word Frequencies in British and American English*. Bergen: The Norwegian Computing Centre for the Humanities.

Kennedy, Graeme. 1998. *An Introduction to Corpus Linguistics*. London and New York: Longman.

Leech, Geoffrey N. and Michael H. Short. 1981. *Style in Fiction. A Linguistic Introduction to English Fictional Prose*. London and New York: Longman.