Reviews

Carol E. Percy, **Charles F. Meyer** and **Ian Lancashire** (eds). *Synchronic corpus linguistics*. Language and Computers: Studies in Practical Linguistics, 16, 1996. Amsterdam Atlanta, GA: Rodopi. vii + 289 pages. ISBN: 90-420-0019-8 (bound), 90-420-027-9 (paper). Reviewed by **Nancy Belmore**, Concordia University.

Synchronic Corpus Linguistics is a well-edited and typographically pleasing selection of papers from the XVIth ICAME conference which was held in Canada in May 1995. The editors present a brief overview of ICAME which would be especially useful for the increasing number of researchers who use corpora but are unfamiliar with the development of corpus linguistics and the role of ICAME in fostering that development. Most of the papers concern themselves with various aspects of corpus development and analysis. Several of these deal with parallel corpora, a reflection of the accelerating development and interest in such corpora. Corpus annotation is less well represented but although there are fewer papers on this topic, their range is impressive.

The four papers on parallel corpora provide a useful compendium of some of the problems in developing parallel corpora, with differing punctuation conventions among the least trivial. Even paragraphs, it would seem, turn out to be unreliable anchor points. The research described here suggests that the prescient techniques for quantifying interlingual distance which Mackey outlined in 1971 could now be fruitfully implemented.

Stig Johansson and Jarle Ebeling describe the English-Norwegian Parallel Corpus (ENPC), exemplifying some of the lexical and syntactic contrasts between the two languages which the Boolean search routines they use reveal. Kay Wikberg uses a subset of this corpus, restricted-length questions of between seven and ten words, as the basis for a contrastive analysis. He shows the potential of such analyses for identifying quite specific features of mis-translations and thus building up a hierarchy of error types as well as revealing syntactic choices in the service of achieving special effects.

Mats Johansson, who uses a sample of texts from the Swedish-English Parallel Corpus to test the common assumption that the constraints on

fronting are less strict in Swedish than in English, suggests that more can be learned about syntactic contrasts by using translations than by studying unrelated texts from each language. Josef Schmied and Hildegard Schäffler note, however, that the phenomenon of 'translationese', which includes deviations from textual norms, ie, instances in which the proportion of particular structures in a translation does not correspond to their usual distribution in the target language, could be misleading. This is why they are developing the 1.5 million-word Chemnitz Corpus which, like the ENPC, will permit studies of both original texts and translated texts.

The eleven other papers on corpus analysis and development describe the use of both classical corpora like the London-Lund Corpus of Spoken English (LLC), newer corpora and corpora which have been specifically designed for the particular project. The uses of these corpora are similarly diversified, ranging from the study of a particular lexical item to a substantial grammar; from purely descriptive studies to particular applications. Sabine Bergler and Sonja Knoll's paper represents an interesting convergence of the techniques of computational linguistics with those of corpus linguistics. They use a newspaper corpus, 79 articles from the Wall Street Journal (28,798 words), to manually identify and categorize noun phrase coreference and to determine which coreference chains, ie, which sets of mutually coreferring noun phrases in a text present the greatest problems in lexical resolution. They find that although most coreferences can be resolved by parsing techniques, a signifcant percentage require complex semantic analysis.

Three papers look at recurrent word combinations, each with a different purpose and each using a different corpus. Antoinette Renouf describes the ACRONYM (The Automatic Collocational Retrieval of Nyms) project which has as its aim the automated identification of useful search terms for text retrieval systems. Nyms are pairs of related words which occur in similar collocational environments. By determining which words share collocates and which words occur significantly often next to each other in a specialized corpus – the ACRONYM corpus is a journalistic corpus of more than 200 million words – an alternative to the standard thesaurus can be developed, one which reflects current usage within a particular domain.

Mats Eeg-Olofsson and Bengt Altenberg have created a database of recurrent word combinations in the LLC. The database contains 37,000 tag sequences of varying length and frequency. One of their purposes is to study what they call 'black holes' or 'gaps', ie, non-recurrent

words or word combinations. Their aim is to determine the constraints on recurrence and to develop contextual rules for tagging those words (14%) which are not part of a recurrent word combination by examining the tags which have been assigned to the 86 per cent which are.

John Milton and Robert Freeman had yet another purpose for looking at recurrent strings or n-grams, defined as combinations of words and marks of punctuation. They hypothesized that the frequency with which L2 writers use collocations is a measure of their proficiency. They present the results of a comparative study of n-gram types in a corpus of 770 L1 examination scripts and in a 750,000-word corpus of L2 (Hong Kong Chinese) examination scripts. They found far fewer differences in proficiency as measured by the use of a variety of collocations among L2 learners at different levels of proficiency than between the most advanced L2 writers and the L1 writers.

Two papers use corpora as the basis for testing previous analyses. Inge de Mönnink evaluates standard descriptions of discontinuous noun phrase modification in terms of the patterns she has found in an examination of the Nijmegen corpus as well as a sample of face-to-face conversations in the ICE-GB corpus. She finds that the descriptions in traditional grammars of the less common constructions are incomplete and even, at times, incorrect. Henk Barkema examined the grammatical flexibility of idioms and other lexicalised expressions in the 20-million-word Birmingham Collection of Texts. Contrary to expectations, the most collocationally limited types were the most flexible, and formulaic types are no less flexible than other types. An ANOVA test showed that the degree of 'compositionality', a measure of the degree to which the meaning of an expression is derivable from its constituent lexical items and the syntactic structure of the resultant expression, was not a reliable predictor of the observed degree of flexibility.

Dieter Mindt's paper describes his inductive grammar of English modals, *An Empirical grammar of the English verb* (1995). Mindt has used an 80-million word collection of texts, primarily unedited fictional texts drawn from a number of existing corpora, to derive an inductive grammar of English modals. The paper describes the new analysis of the English verb phrase which resulted. The book gives an account of the form and meaning of each modal based on statistics from the corpus. It then identifies the modals associated with the expression of particular meanings as well as the grammatical contexts in which the modals occur.

Several papers look at language variation which is attributable to the

extra-linguistic contexts in which a text occurs. Juhani Norri and Merja Kytö describe the compilation of the Tampere Corpus of English for Specific Purposes, a stratified corpus for investigating linguistic variation in scientific texts. The final corpus will include ten 3,000-word texts at each of four levels of technicality, ie, 40 texts per field, drawn from ten different fields, for a total of 1.2-million words. A pilot study of medical and biological texts found a significant relation between level of technicality and the frequency of referential pronouns as well as the frequency of everyday words vs more formal or technical words. Noting the frequency of words referring to fighting or warfare in the medical texts, the authors suggest that entire lexical domains may vary in frequency as a function of level of technicality and/or field.

Magnus Ljung tests the conflicting claims that the occurrence of non-finite and verbless adverbial clauses is relatively even across genres or, alternatively, largely restricted to more formal texts. In an examination of texts from ICE-GB, a 200,000-word corpus of academic texts and a 600,000-word American and British newspaper corpus, he found that certain genres more fully exploit the potential use of such clauses, with conversations predictably the least likely to do so and British science texts, the most likely.

Bas Aarts investigates the use of the word *simply* in the different text categories in ICE-GB. His findings support Biber's (1988) demonstration that shared textual dimensions can outweigh genre in predicting the occurrence of particular grammatical and lexical features. Aarts identifies 'persuasive function' as an important predictor variable thereby recalling Biber's textual dimension 'overt expression of persuasion'. The ICE categories in which *simply* occurs with high frequency all have this rhetorical function in common while the opposite is generally true of the low frequency categories.

Anna-Brita Stenström and Gisle Andersen use the Bergen Corpus of London Teenage Language (COLT) and the LLC to study variation as a function of age. They focus on two forms, *cos* and *innit*. *Cos* occurs in both corpora but unlike the adult speakers in the LLC, who prefer *because*, teenagers prefer *cos*. They use *cos* as a non-subordinator fulfilling a variety of pragmatic functions far more often than as a subordinator. *Innit* does not occur at all in the LLC and it too fulfills numerous pragmatic functions in the COLT samples they studied. They conclude that the use of the two forms as pragmatic particles is a linguistic innovation specific to the speech of teenagers.

The final group of articles is concerned with the development and

evaluation of corpus annotation. Alex Chengyu Fang has used that part of ICE-GB which was tagged and parsed by the TOSCA parser developed at Nijmegen University to arrive at a modified tagger (AUTASYS) and parser called the Survey Parser. Both systems are described in considerable detail as well as a comparison of the Survey Parser with the XTAG and Alvey Natural Language Tools parsers. The Survey Parser yielded results which compared favourably with the other two systems.

A. M. Wallington, M. D. Dennis and G. R. Sampson take a completely different approach to parsing. Their paper describes the special features of APRIL3, which uses the stochastic optimization technique of simulated annealing to search out the most plausible grammatical analysis of an input string. The statistical model which it uses to evaluate plausibility is derived automatically from the analyzed SUSANNE corpus.

Tim Willis describes his experience in using the TOSCA parser with the Lancaster Corpus of Spoken English. He found that user experience is a key factor in achieving a high success rate and he identifies some of the cases in which the parser fails. He concludes that TOSCA is a very powerful system whose main problems are the non-standard and debatable structures which linguists themselves have trouble classifying.

Graeme Kennedy outlines the inaccuracies which may arise when using a tagger designed for one dialect with another dialect. While a probabilistic system like CLAWS, which was designed to tag and parse British English, may work remarkably well on another dialect, there are limitations. Kennedy illustrates this by looking at how CLAWS 1 tags the word *once* in the Welling Corpus of New Zealand English. Even where tagging is accurate, sense disambiguation may be required since the meanings associated with identically (and correctly) tagged words may differ across dialects. For all these reasons, some manual analysis will always be required and comparative studies must be based on using identical taggers and the same post-editing techniques.

In the final paper in the volume, John M. Kirk considers the problem of designing an adequate transcription and annotation scheme for spoken corpora. Violations of strict linear progression, such as interruptions and overlapping speech, must be modeled in the transcription, and a way must be found to annotate the transcription without impairing readability. He presents his own proposal for an annotated transcription which indicates the discourse functions of the utterances and facilitates a quantitative and qualitative analysis of those lexical items which only occur in speech, the so-called ‘discourse items’ or ‘interaction signals’.

References

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Mackey, William Francis. 1971. *La distance interlinguistique*. Quebec: Les Presses de l'Université Laval.

Mindt, Dieter. 1995. *An Empirical grammar of the English verb*. Berlin: Cornelsen Verlag.

Tony McEnery and **Andrew Wilson**. *Corpus Linguistics*. Edinburgh: Edinburgh University Press, 1996. ISBN 0-7486-0808-7 (hardback); ISBN 0-7486-0482-0 (paperback). Reviewed by **Charles F. Meyer**, University of Massachusetts at Boston.

The publication of *Corpus Linguistics* is noteworthy: as the first volume in the new series 'Edinburgh Textbooks in Empirical Linguistics', this textbook reflects not only the increasing importance that empirically-based studies of language are coming to play in linguistics but the prominent role that corpus linguistics has assumed among the many different empirically-based approaches to language study. In *Corpus Linguistics*, McEnery and Wilson (hereafter MW) very clearly introduce the field of corpus linguistics to students, providing a very effective overview of the key linguistic and computational issues that corpus linguists have to address as they create corpora and conduct analyses of them.

Corpus Linguistics is divided into seven chapters that focus on a number of topical issues in corpus linguistics, issues ranging from the theoretical underpinnings of corpus linguistics to the various annotation schemes that have been developed to tag and parse corpora, the quantitative research methods used to analyze corpora, the types of linguistic studies that have been carried out on corpora, and the contributions that computational linguistics has made to the creation and analysis of corpora. Each of these topics is approached in a clear and readable format that will make this text valuable not just to students but to specialists in other areas of linguistics interested in obtaining information about corpus linguistics.

After noting in the opening chapter (‘Early Corpus Linguistics and the Chomskyan Revolution’) that corpus linguistics is more a methodology (a way of approaching language study) than a sub-discipline in linguistics, MW continue with a discussion of the methodological assumptions that characterize corpus linguistics and distinguish it from Chomskyan approaches to language study. They note the difference between rationalist and empiricist approaches to language study, and detail the classic Chomskyan arguments that have been leveled over the years against empiricist studies of language. Because corpora contain data reflecting ‘performance’, they are of little value in studying ‘competence’, the most important area for linguists to study. In addition, ‘corpora are “skewed”’ (p 8), in the sense that they do not contain all of the possible structures that exist in a language. These objections led Chomsky to value introspection as the best way of describing a language, and to reject descriptions of actual language use based on analyses of corpora.

Although MW acknowledge some validity to Chomsky’s objections to corpus analyses, they counter these objections with a number of arguments in favor of corpus linguistics. A corpus, for instance, can be used to verify introspective judgments, and to overcome the problem of basing grammatical arguments on ‘artificial data’ (p 12). Moreover, corpora can provide important information on the frequency of grammatical constructions, and the sophisticated software developed to analyze corpora can give the linguist access to much important information on grammatical structure present in corpora that have been tagged and parsed.

Although Chapter 2 (‘What is a corpus and what is in it?’) purports to describe what a corpus is, it is primarily a chapter about what corpora look like—specifically the annotation schemes that have been developed to tag and parse them. MW only briefly discuss the issues one must confront when creating a corpus (eg the size of the corpus), and while they discuss many methodological concerns throughout the book, it would have been desirable to have grouped these issues together in a single chapter and to have discussed how the representativeness of a corpus is influenced by such variables as its length, the genres it contains, and the types of individuals whose speech and writing are included in the corpus.

The strength of Chapter 2 is its discussion of annotation schemes, which is detailed and very well illustrated. MW provide a very clear overview of the TEI (Text Encoding Initiative), illustrating how the various tags developed by TEI can be used to create ‘headers’ (in which information about authors/speakers, titles, dates of publication, etc can

be recorded) and to mark up texts themselves with information on paragraph boundaries, type faces, and so forth. The remainder of the chapter focuses on the various schemes that have been developed to annotate linguistic information in corpora. MW first compare tagging schemes from corpora as diverse as the British National Corpus and the CRATER Corpus of Spanish, and then describe the process of developing the CLAWS tagging schemes at Lancaster University. The chapter concludes with a discussion of parsing schemes and of how corpora can be annotated with markup revealing their semantic, discoursal, and prosodic structure.

Chapter 3 ('Quantitative data') discusses the importance of using quantitative research methods to analyze corpora. MW first distinguish qualitative from quantitative research methods, and make the very important point that the linguistic claims one makes about a corpus depend crucially upon whether the corpus being analyzed is valid and representative; that is, has been created in a manner that allows the analyst to make general claims about, for instance, the genres represented in the corpus. MW then describe the major kinds of statistical analyses that can be performed on corpora. The difficulty of a chapter of this type is that statistics is such a vast area that it is hard to determine precisely how much detail needs to be provided. But the level of detail in this chapter is most appropriate, and there is much useful information provided on how corpora can be statistically analyzed — from methods as basic as frequency counts to those as sophisticated as factor analysis and loglinear analysis (as done with programs such as VARBRUL).

Chapter 4 ('The use of corpora in language studies') surveys the kinds of empirical linguistic analyses that corpora can be used to conduct. MW open the chapter with a discussion stressing the importance of empirical studies of language, noting that they 'enable the linguist to make statements which are objective and based on language as it really is rather than statements which are subjective and based upon the individual's own internalised cognitive perception of the language' (p 87). This statement is very convincingly supported in the remainder of the chapter, which contains a very good discussion of how corpora can be used to study language at all levels of linguistic structure (eg phonetics/phonology, syntax, and semantics) and from many different theoretical perspectives (eg pragmatics, sociolinguistics, and discourse study). As each of these areas are described, MW include descriptions of previous studies conducted in the areas to effectively illustrate how work in the area is conducted and has yielded important information.

The first four chapters of *Corpus Linguistics* are concerned with issues relevant to linguists using corpora to carry out purely linguistic studies. Chapter 5 ('Corpora and computational linguistics') moves to an allied discipline, natural language processing (NLP), and discusses issues such as tagging and parsing from a more computational perspective. Although linguists who use corpora for grammatical analysis may not have an immediate interest in NLP, the research in this area has led directly to improvements in recent years of taggers and parsers — software responsible for annotating corpora and making it easier for linguists to extract information from them. The discussion in this chapter is brief, but does an excellent job of summarizing the theoretical issues underlying the development of taggers and parsers and the role that they play in areas such as lexicography and machine translation.

Chapter 6 ('A case study: sublanguages') draws upon much of the information presented in the previous chapters to carry out a sample grammatical analysis of three corpora from three distinct genres: a series of IBM manuals from the IBM Corpus, transcriptions of Canadian parliamentary speeches found in the Hansard Corpus, and fiction from the APHB (American Printing House for the Blind) Corpus. MW analyze these three corpora to advance the hypothesis that the language of the IBM Corpus is a 'sublanguage'; that is, 'a version of a natural language which does not display all of the creativity of that natural language[and which] will show a high degree of **closure** [emphasis in original] at various levels of description (p 148). Before pursuing this hypothesis, MW discuss the importance of evaluating a prospective corpus to determine whether the genres it contains are appropriate for the analysis being conducted, and whether the manner in which the corpus has been annotated will allow for the retrieval of the grammatical information desired. MW conclude that the corpora they have chosen are appropriate for their study of sublanguages, and they then conduct analyses to determine the degree of lexical closure, part-of-speech closure, and parsing closure that exists in each of the corpora. In general, this analysis verified MW's hypothesis and demonstrated that the IBM Corpus (in comparison with the other corpora) is a more 'restricted genre' and contains fewer different types of words and sentence-types (though the words it contained corresponded to more parts of speech than the words in the other corpora did).

The final chapter ('Where to now?') nicely rounds out the book with a discussion of issues that corpus linguists will need to address in the future corpora that they develop, a series of 'pressures' to increase the

length of corpora, to make them conform to industry as well as academic standards, and to have corpora draw upon evolving computer technologies in their creation, such as the many multi-media currently being developed. This chapter provides a fitting conclusion to a text that provides a very perceptive overview of the field of corpus linguistics that will be a good choice for use in any introductory course on corpus linguistics.