Sampson, Geoffrey and Diana McCarthy. *Corpus Linguistics: Readings in a Widening Discipline*. London, New York: Continuum International Publishing Group Ltd. 2005.

The book is published in the Open Linguistics Series, which offers other high–quality linguistic titles. The subject of corpus linguistics has been a propulsive one in the last half century. The present volume gives a cross–section in the way of a chronology of the development of corpus linguistics. In addition to the *Introduction*, the book contains 42 articles or chapters, formerly published elsewhere, spanning the period from 1952 to 2002, each with an editorial introduction which explains the purpose of its inclusion in the book. The majority of the editorial comments are of a nature that can be agreed with, although a few could be further supplemented. There is a hiatus in the selection of publications between 1952 and 1965 for obvious reasons, and then again between 1971 and 1986, which was not as bleak a period as may seem. The book otherwise traces exceptionally well the directions in which corpus linguistics has been developing.

In the *Introduction* (pp. 1–8), in addition to the statement "people have studied languages via corpora for a long time" further information on the history of the techniques used in corpus linguistics might have been given Concordancing and word counting, as major techniques of corpus linguistics, have a much longer history than the discipline itself. In the pre–computer era, concordances and word frequency counts were prepared for major works such as the Bible (in several languages), the Qur'an, and Shakespeare. Word counts were used for language teaching purposes (Thorndike, West) and, as mentioned in the Introduction, for stenography (Kaedig). Some authors mention old Greek, Indian, Hebrew, Latin and Arab works, serving as invaluable information for readers. The history of corpora dates much farther back in history than Johnson and the 18th century, so there is a feeling that a more elaborate historic overview of corpus linguistics is lacking. However, it is true that it started to flourish in the mid 1960s due to the development and availability of computers, but the techniques had been used much earlier.

Chapters 2 to 43 include pieces which may be tentatively grouped into four broad areas:

a) theoretical and review: From *The Structure of English* (1952) Charles Carpenter Fries; *Treebank grammars* (1996) Eugene Charniak; *Reflections of a dendrographer* (1999) Geoffrey Sampson

b) linguistic level analyses and research: *On the distribution of noun–phrase types in English clause–structure* (1971) F. A. M. Aarts, *Predicting* *text segmentation into tone units* (1986) Bengt Altenberg; *Typicality and meaning potentials* (1986) Patrick Hanks; *Historical drift in three English genres* (1987) Douglas Biber and Edward Finegan; *Cleft and pseudo–cleft constructions in English spoken and written discourse* (1987) Peter C. Collins; *A point of verb syntax in south–western British English: an analysis of a dialect continuum* (1991) Ossi Ihalainen; *On the history of that /zero as object clause links in English* (1991) Matti Rissanen; *Structural ambiguity and lexical relations* (1993) Donald Hindle and Mats Rooth; *Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies* (1993) William Louw; *Conflict talk: a comparison of the verbal disputes between adolescent females in two corpora* (1996) Ingrid Kristine Hasund and Anna–Brita Stenström; *Linguistic and interactional features of Internet Relay Chat* (1996) Christopher C. Werry; *Distinguishing systems and distinguishing senses: new evaluation methods for word–sense disambiguation* (1997) Philip Resnik and David Yarowsky; *Qualification and certainty in L1 and L2 students' writing* (1997) Kenneth Hyland and John Milton; *Analysing and predicting patterns of DAMSL utterance tags* (1998) Mark G. Core; *Assessing claims about language use with corpus data – swearing and abuse* (1998) Anthony McEnery et al.; *The syntax of disfluency in spontaneous spoken language* (1998) David McKelvie; *The use of large text corpora for evaluating text–to–speech systems* (1998) Louis C. W. Pols et al.; *Intonational variation in the British Isles* (2002) Esther Grabe and Brechtje Post.

c) corpus creation, corpus methodology and use: *A standard corpus of edited present–day American English* (1965) W. Nelson Francis; *Corpus creation* (1987) John Sinclair; *Using corpus data in the Swedish Academy grammar* (1991) Staffan Hellberg; *Encoding the British National Corpus* (1992) Gavin Burnage and Dominic Dunlop; *Computer corpora – what do they tell us about culture?* (1992) Geoffrey Leech and Roger Fallon; *Representativeness in corpus design* (1992) Douglas Biber; *A corpus–driven approach to grammar: principles, methods, and examples* (1993) Gill Francis; *Building a large annotated corpus of English*: *the Penn Treebank* (1993) Mitchell P. Marcus et al.; *Automatically extracting collocations from corpora for language learning* (1994) Kenji Kita et al.; *Developing and evaluating a probabilistic LR parser of part–of–speech and punctuation labels* (1995) E. J. Briscoe and J. A. Carroll; *English corpus linguistics and the foreign–language teaching syllabus* (1996) Dieter Mindt; *Why a Fiji corpus?* (1996) Jan Tent and France Mugler; *The Prague Dependency Treebank: how much of the underlying syntactic structure can be tagged automatically*? (1999) Alena Böhmová and Eva Hajicová; *A generic approach to software support for linguistic annotation using XML* (2000) Jean Carletta et al.; *Semi–automatic tagging of intonation in French spoken corpora* (2001) Estelle Campione and Jean Véronis; *Europe's ignored languages* (2001) Anthony McEnery; *Web as corpus* (2001) Adam Kilgarriff

d) corpus statistics and quantitative studies: *What is wrong with adding one?* (1989) William Gale and Kenneth Church; *A statistical approach to machine translation* (1990) Peter F. Brown et al.; *Data–oriented language processing: an overview* (1996) L. W. M. Bod and R. J. H. Scha; *Assessing agreement on classification tasks: the kappa statistic* (1996) Jean Carletta

In the introductory note to Chapter 3 (*A standard corpus of edited present–day American English*), which presents a description of a seminal electronic corpus in the history of linguistics published in 1964, the editors point out: "The late 1960s and early 1970s were the period when hostility to empirical and quantitative methods within the discipline of linguistics was at its peak, with the consequence that the Brown Corpus was relatively neglected during the first decade of its existence" (p. 27). However, it should be noted, for the sake of truth, that my own interest in corpora goes back to 1971 when I worked at the Institute of Linguistics at the Faculty of Philosophy, University of Zagreb, which was in 1967 granted permission to use (half of) the Brown Corpus to conduct the Serbo–Croatian–English Contrastive Project (The Faculty of Philosophy, pp. 159–160). The Project extensively cooperated with the Centre for Applied Linguistics, Washington, D. C. For the purpose of the Project, "contrastive concordances", i. e. English originals and their Croatian translations, were used, so the Zagreb version of the Brown Corpus was hence bilingual, and all the words were accompanied by assigned codes. A numerical coding system was used for part–of–speech categories and grammatical functions in order to enable morphological and syntactic elements to be retrieved. At that time, punched cards were used, and the sentences to be analyzed contrastively were printed on paper slips of the size of punched cards, which were then stored in boxes, and later taken out and paired as they were needed for the analysis. The results of the research were published in a series of publications: A. *Studies*, B. *Reports* and C. *Pedagogical Materials*, and the major publication *Chapters in Contrastive Grammar of Serbo–Croatian and English*. In the same period, numerous other contrastive projects were conducted in Europe, and probably elsewhere, such as English–German, the Poznan Polish–English Project, English–Hungarian, English–Chech, Romanian–English, English–Spanish and numerous other projects and studies, some of which were based on corpora (Filipović, 1971). So, mention made of these could have filled the 1971–1986 hiatus.

Some of the fifty years under review are represented with more than one selected piece, so 1996 is represented with seven articles (which are not all of equal importance and use, for example, Ch. 30), and then come 1993 and 1998 with four. Some chapters could easily have been omitted without damaging the overall fine content of the book (e. g. Ch. 40). A chapter selected, for example, from the work of Sue Atkins or some other lexicographer regarding corpus lexicography might also have been added.

The *Bibliography* (pp. 483–509) is a merged list of the references from the originals reprinted, within which a list of conventional abbreviations for the proceedings of frequently cited conference series is given. This is followed by a

URL List (pp. 509–510) containing very useful web addresses (the problem is that some of them are no longer active). The last part is the *Index* (pp. 511–524) where names and technical terms are given together, which makes it less easy to use. In the following editions, it might be preferable to give these separately.

Although it is hard to say that anything should be added to a book of over 500 pages, some materials on the use of corpora in teaching and in the production of teaching materials might have been added (for example, the very broad field of compilations of minimal vocabularies was not mentioned) and, as previously indicated, the whole field of contrastive linguistics.

The comments given do not lessen the value of the reviewed volume. They seek merely to contribute to its completeness, since it is a volume to be highly recommended as a reader and as a reference book.

References

Corréard, Marie–Hélène. (ed.) 2002. *Lexicography and Natural Language Processing*. EURALEX.

Filipović, Rudolf. (ed.) 1971. *Zagreb Conference on Contrastive Projects*. Zagreb: Faculty of Philosophy, Institute of Linguistics

Muhvić–Dimanovski, Vesna. *Institute of Linguistics*. In: *The Faculty of Philosophy at the University of Zagreb*. Zagreb: FF Press. 2004, pp. 159–163.

Milica Gačić