Abstracts

Guy Aston and **Lou Burnard**. *The BNC Handbook: Exploring the British National Corpus with SARA.* Edinburgh: Edinburgh University Press, 1998. 256 pp. ISBN 0-7486-1054-5 (cased), 0-7486-1055-3 (paperback). Abstracted by **Ylva Berglund**, Uppsala University.

The British National Corpus (BNC) is a 100 million word corpus of contemporary written and spoken British English. This is an important source of information for linguists, lexicographers, language engineers and others interested in the English language. A special software system, SARA (SGML-Aware Retrieval Application) was developed to enable easy access to the BNC data. The software is distributed with the corpus, and also used on the on-line network service available. The present volume provides an introduction to the British National Corpus (BNC) and the SARA software system.

The book consists of three parts. In the first part, the reader is given a comprehensive and up-to-date introduction to corpus linguistics and corpora in general and the BNC in particular. Issues dealt with include potential uses of corpora, previous corpus-based research, corpus design and corpus annotation.

The second, and most extensive, part of the book contains a guide to using the BNC with SARA. In ten exercises, the reader is presented with various research questions and it is shown, in an easy to follow, step-by-step manner, how the questions can be solved using the BNC and SARA. Among the topics covered are current usage, collocation patterns, text type variation, sociolinguistic variables, linguistic ambiguity, syntactic variation, and pragmatics. The ten exercises are structured similarly. First the research problem is presented and the features of the software to be used listed. Then follows the step-by-step procedure of solving the problem, with comments on each step. Each section is concluded with a discussion of the exercise and with suggestions for further work. It is recommended that the exercises be worked through in the sequence they are presented since the features of the corpus and software are introduced progressively. However, frequent references to other sections and the detailed description of the procedures make it possible to use the exercises for reference purposes, too.

The third section of the book is the Reference Guide. Apart from a number of code tables and a definition of the SGML Listing format (used for saving query results to a file), the section contains a 'Quick reference guide to the SARA client'. This guide provides a summary of the facilities available in the SARA program. The features of the software are presented as they appear on the menu bar in the program, thereby constituting a valuable supplement to the previous, task-based description.

Geoff Barnbrook, *Language and Computers: A Practical Introduction to the Computer Analysis of Language.* Edinburgh: Edinburgh University Press, 1996. 209 pp. ISBN 0-7486-0848-6 (cased), 0-7486-0785-4 (paperback). Abstracted by **Ylva Berglund**, Uppsala University.

In this volume, Geoff Barnbrook gives an introduction to the use of computers in the analysis of language. Starting out by asking 'Why use a computer?', he compares traditional manual methods to computer-based approaches. By way of illustration, authentic examples are provided from actual studies performed, showing what computers can and cannot do. It is shown that although much can be done more quickly and cheaply (in terms of time and money spent) with computers, some things are still better done in the traditional, manual way. Before deciding whether to use computers, it is therefore important to make careful calculations of the costs involved. The book provides a check-list to help make that kind of calculation.

Moving on from the basic question about the use of computers, the author discusses how to capture the data in computer-readable form. Sources of readily available text are identified and methods of making text machine-readable are illustrated. Other issues concerning the capture and use of data are discussed, and potential problems are presented with suggestions as to how they can be solved.

Three chapters (3-5) then discuss various ways of dealing with a machine-readable text. The use of frequency lists, concordances and collocation analysis are presented and analysed. Possible problems are identified and solutions are suggested. Frequent illustrations from various studies are provided throughout.

Having dealt primarily with data consisting of plain text, the author moves on to discussing the usefulness of detailed linguistic information (syntactic tagging, parsing and semantic disambiguation). Approaches making such information available to the computer are illustrated. It is shown that manual annotation is time-consuming but that it can be difficult to achieve the necessary accuracy with fully automated processes.

The use of computers in language study is put into a broader perspective in the chapter on applications of natural language processing (Chapter 7). It is shown how computerised language analysis is involved in areas of word processing, lexicography, data retrieval, computer-assisted language learning, expert systems, and machine translation. In the concluding chapter of the book (Chapter 8), some case studies from 'real research projects' (p 149) are presented, illustrating the kind of research tasks where computers can be used and the kind of problems that can occur.

The book concludes with a glossary and four appendices, ie, (1) Programming languages for language programming, (2) awk: a very brief introduction, (3) Detailed programming examples, and (4) Suggestions for the exercises.

Roger Garside, Geoffrey Leech and **Anthony McEnery** (eds*). Corpus Annotation: Linguistic Information from Computer Text Corpora*. New York: Addison Wesley Longman Limited, 1997. 281 pp. ISBN 0-582-29837-7. Abstracted by **Ylva Berglund**, Uppsala University.

In this volume the reader is taken through some of the latest research in the field of corpus annotation. In 17 chapters, a number of authors, many of them responsible for more than one paper, introduce, describe and illustrate various areas of corpus annotation. Though the chapters can be read as separate papers on particular aspects of corpus annotations, many of them are introduced briefly with reference to preceding chapters, their content thereby being placed within the larger framework of corpus annotation.

After an introduction to corpora and corpus annotation systems, written by Geoffrey Leech, grammatical word tagging, syntactic, semantic, and discourse annotation are treated in separate chapters (Chapters 2-5), whilst the less developed areas of prosodic, pragmatic, and stylistic annotation are presented together in Chapter 6. The chapters are similar

in structure, with an introduction to the annotation system in question, followed by a presentation of it in greater detail, often with illustrations drawn from existing corpora and with examples of how the annotated corpora have been and can be used in linguistic research.

In the following chapters (Chapters 7–10), the reader is introduced to the more practical aspects of word class tagging through descriptions of the process of annotating text corpora. First the CLAWS4 tagger is presented against a background of probabilistic and rule-based taggers. The following two chapters explain how the performance of the CLAWS tagger has been enhanced (Chapter 9) and how the Template Tagger, among other things, has been developed as a complement to CLAWS (Chapter 8). As an illustration of the tagging of other languages than English, Chapter 10 describes how the probabilistic Xerox Tagger was retargeted to apply to Spanish.

In Chapter 11, the authors deal with methods of providing a corpus with syntactic annotation, while Chapter 12 examines some tools used for higher-level annotation, such as introduced in Chapters 4 and 5. Manual and automatic procedures are presented, and it is shown that a combination of the two approaches seems to be used in most applications. Chapter 13 examines corpus/annotation tools needed for creation and exploitation of annotated corpora, while the following chapter (14) focuses on the Cytor program, showing how it has been successfully used for teaching basic grammatical word classes. Another example of how annotated corpora can be used is given in Chapter 15, where the use of multi-lingual (parallel) corpora for terminology extraction is discussed.

The last two chapters of the book deal with some issues of general interest for various kinds of annotation. Chapter 16 looks at the problem of finding a standard for annotation. It is proposed that a set of recommendations and guidelines on different levels (obligatory, recommended, and, optional) be used instead of rigid standards. In Chapter 17 it is shown that human post-editors can improve the accuracy and consistency of automatically annotated text. The book concludes with three appendices providing a list of World Wide Web and email addresses, a glossary of abbreviations and acronyms, and the C7 and C5 tagsets referred to in the book.

Michael Stubbs. *Text and Corpus Analysis: Computer-assisted Studies of Language and Culture.* Oxford: Blackwell Publishers Ltd. 1996. 267 pp. ISBN 0-631-19511-4 (cased), 0-631-19512-2 (paberback). Abstracted by **Ylva Berglund**, Uppsala University.

The present volume places the use of corpora and computational methods firmly within the (socio)linguistic framework. Great weight is given to the importance and usefulness of authentic data and computational methods. The main focus is, however, on the linguistic analyses of the data and what these can tell us about the language, its users and their society.

Against the background of British linguistics, Stubbs shows how computer-assisted methods can be used to analyse texts. A major topic of concern is the importance of using 'real data': ie actual, attested, authentic examples/texts which have been produced without the intervention of the analysing linguist. The main question discussed is: 'How can analysis of the patterns of words and grammar in a text contribute to an understanding of the meaning of the text?' (p 3).

The book consists of two parts. The first, theoretical part, begins with a presentation and discussion of some central concepts, such as the concepts of text, text type and genre. Then follows a discussion of ideas that have developed in British linguistics from the 1930s to 1990s, starting with Firth and further developed by Halliday and Sinclair. Some central principles are outlined, defining linguistics as a social science and applied science, emphasising the importance of studying authentic data, preferably in whole texts. It is claimed that form and meaning cannot be separated; nor can lexis and grammar be seen as independent concepts. These principles are developed and discussed with illustrations from previous research and in contrast with the ideas presented by, for example, Noam Chomsky. Some 'broader sociological points' (p 51) are developed and the concept of institutional linguistics is presented. It is maintained that an analysis of language requires a combination of studying authentic data (corpus analysis) and looking at the framework within which it occurs (institutional analysis).

The second part of the book presents analyses of texts and corpora. It is shown how the study of selected linguistic features can reveal significant social and cultural characteristics. Among the features treated in the analyses are the ways ideological (sexist) positions are conveyed by individual words and patterns of vocabulary and grammar. It is also

discussed, for example, whether lexical and grammatical features of a judge's style can influence a jury and how analysis of syntactic and semantic features in different texts can reveal differences based on ideological stances. It is concluded that '...texts, spoken or written, are one of the empirical, material bases of society, and computer-assisted analysis reveals some of the everyday routines which reproduce social institutions' (p 237).

Anne Wichmann, Steven Fligelstone, Tony McEnery, and **Gerry Knowles** (eds). *Teaching and Language Corpora.* New York: Addison Wesley Longman Limited, 1997. 343 pp. ISBN 0-582-27609-8. Abstracted by **Ylva Berglund**, Uppsala University.

This book is aimed at language teachers in higher education and applied linguists with an interest in learner-centred approaches. The volume contains a varied range of papers dealing with different aspects of teaching and language corpora. Most of the papers were first presented at the *Teaching and Language Corpora* conference (TALC94) in 1994.

Apart from the introductory chapter, where Geoffrey Leech discusses 'the convergence of language teaching and language research, through the link of corpus-based methods' (p 1), the volume consists of four sections, each dealing with a particular aspect of teaching and language corpora.

Section A, 'Why Use Corpora', contains three chapters where the role and potential of corpora used in language teaching are discussed. The nature of language itself is dealt with as well as the learning process and the role of the teacher. Among other issues, the authors compare the traditional language teaching method, with reliance on existing descriptions of the language, with the corpus-based approach.

The seven chapters in section B, 'Teaching Languages', evolve around issues relating to the teaching of a language, here illustrated with examples from English, German, and Welsh. Practical aspects of how corpora have been used in language teaching are discussed, and various corpora and corpus tools, such as concordancing tools, are dealt with. In connection to this, attention is also paid to the role of the teacher and the student.

Section C, ‘Teaching Linguistics’, consists of six chapters where the authors explore how corpora can be used in the teaching of linguistics. The areas dealt with include micro- and macro linguistics, child language acquisition, the diachronic study of English, the teaching of prosody, as well as the investigation of style and the role of corpora in critical literary appreciation. To a large extent the papers report on experiences drawn from actual courses taught, with illustrations showing the benefits to teachers and students of using a corpus-based approach.

In section D, ‘Practical Perspectives’, four chapters deal with different practical matters relating to the use of corpora in teaching. Starting with how to teach teachers to use corpora in research and teaching, the authors then move on to describing the building of a corpus from readily available textual resources and the creation of corpora in Greek and Cyrillic alphabets. The last chapter outlines some main areas that may need consideration in developing a computing infrastructure.

The book contains a small reference section (Appendices 1-5), where the reader can find a brief guide to further investigation, a list of sources of information and electronic texts, a list of corpora and software mentioned in the book, and a glossary of computing terms.

Throughout the book the authors argue in favour of an increased use of corpora in language teaching and teaching about languages. The arguments are supported by illustrations from corpus-based methodology, providing advice for teachers on how to use corpora in teaching as well as examples of results obtained when using this approach. Certain issues are dealt with in more than one section. One such issue is the benefit of using newspaper corpora. Another is the role of the teacher and the student. It is argued more than once that, when corpora are used in language teaching, the students are made to assume the more active role of the researcher, making discoveries about the language. A natural consequence of this is that the role of the teacher is changed from that of being the sole provider of knowledge to that of someone guiding and supporting the students in their quest for knowledge.