Reviews

Paul Baker, Andrew Hardie and **Tony McEnery**. *A glossary of corpus linguistics*. Edinburgh: Edinburgh University Press, 2006. 187 pp. ISBN 10-0-7486-2403 1 (hardback), ISBN 10-0-7486-2018 4 (paperback). Reviewed by **Helena Raumolin-Brunberg**, University of Helsinki.

The increasing availability of electronic corpora means that every year more and more new people enter the field of corpus linguistics. I think it is especially these people that can benefit from the *Glossary of corpus linguistics*, compiled by three prominent corpus linguists from the University of Lancaster. This is not to say, of course, that the entries would not be helpful even for more advanced corpus linguists.

The book contains nearly 500 entries, organized alphabetically, and a separate list of acronyms used in corpus linguistics. The entries are short, rarely exceeding 100 words, and this means, of course, that the information is not exhaustive. References to further sources of information are available for some, but not all entries. Website addresses have only been given for "some organisations, groups, corpora or software where we feel that the site is unlikely to close down or move" (p. 1).

The authors have not expressed their principles for the inclusion of topics, but the blurb on the back cover lists six areas of focus: (1) important corpora, (2) key technical terms, (3) key linguistic terms relevant to corpus-based research, (4) key statistical measures used in corpus linguistics, (5) key computer programme/retrieval systems used in the construction and exploitation of corpora, and (6) standards applied within the field of corpus linguistics.

This list illustrates the challenge the compilers must have faced during their project. Apart from the self-evident issues, such as major corpora and the technical terminology of the field, they have had to include terms from linguistics, statistics, and computer science. Drawing borderlines cannot have been easy. The result is a very versatile collection of topics, which can be illustrated by the entries under the letter N:

named entity recognition
national corpus
natural language processing (NLP)
neologisms
Network of Early Eighteenth-Century English Texts (NEET)
Newcastle Electronic Corpus of Tyneside English (NECTE)
Newdigate Letters
n-gram
N-gram statistics package
Nijmegen Corpus
Nijmegen Linguistic Database
non-parametric test
non-standard corpus
normal distribution
Northern Ireland Transcribed Corpus of Speech (NITCS)
Nota Bene Discourse Annotation Tool
Notetab Light
Nptool

On the whole, there seems to be a good balance between the different components. My general impression is that the explanations are more thorough in the areas of technical terminology, annotation, statistics and computer science than linguistics. At times, one wonders if some of the terms could have been left out on the assumption that the readers will have become familiar with them in their specific fields of study, for example *conversation analysis*, *introspection*, *lexeme*, and *postmodification*. This would have given more room for a deeper discussion of the central topics and especially for references for further reading.

Although the entries include corpora of several of the world's languages, such as Chinese, Japanese, Korean, French, Greek, German, and Spanish, the focus is clearly on English, with corpora from various continents and time periods. This bias on English may diminish the usefulness of the book for those who investigate other languages, although the methodological entries are, of course, valid irrespective of the language to be studied.

As regards the methodological tools and search programs, the contact information is sufficient for most of the entries, but in some cases it is missing. It is, for instance, good for the reader to know what *Varbrul programs* are for, but giving the names of the creators of these tools or, better still, a website to contact would have been helpful.

The biggest problem for this kind of book is the rapid expansion and change in the field of study. Change is evident in all of the six areas the volume covers, but it hits hardest in the presentation of important corpora. Although the book introduces about one hundred corpora, it seems that some new ones are missing. For example, I would have liked to see entries for the *Corpus of English Dialogues* and the *Penn-Helsinki Parsed Corpus of Early Modern English*. As far as tools are concerned, I think the *Corpus Presenter* would also have been worth an introduction.

In addition, I would have welcomed information on the more recent development of several of the corpora included. As an illustration I will consider the corpus project that I have been involved in since 1993, the *Corpus of Early English Correpondence* (CEEC). The *Glossary* gives accurate information about the corpus on the basis of an article from 1998 (Keränen; not in the bibliography), but nothing is mentioned concerning its development in the 2000s. In other words, the extension to cover the period 1680-1800 and the parsing of the original corpus in a joint project between the University of Helsinki and the University of York are not discussed. These processes have been reported in several publications and on the corpus website.

It is not only in the presentation of some of the corpora that the information seems somewhat outdated. For instance, the entry on *variation* contains three studies as illustrations; two are from 1992 and the third from 1998. This is a field where a great deal of research has been carried out in recent years, and I would have liked to see examples from the current decade.

It may be that the interval between the compilation of the *Glossary* and its publication has simply been too long for the newest reports to be taken into account. This possibility leads me to the question of the mode of publication of this type of material in a rapidly changing field like corpus linguistics. I think glossaries and dictionaries dealing with science-specific data are the type of material for which electronic publication on the internet would be more suitable than the printed word. Electronic publication would allow regular updating of the material, and the newest developments could reach users with less delay. This, of course, assumes that the necessary resources would be made available.

The above comments are not to say that the book is not a very valuable source of information about corpus linguistics, a branch of study that combines many different disciplines and is therefore short on comprehensive sources of information. I think the book also provides good material for teaching, and the list of acronyms is certainly welcome to every corpus linguist. In future, however, I hope to be able to find updated versions in an electronic format.

Roberta Facchinetti (ed.). *Corpus linguistics 25 years on* (Language and Computers 62). Amsterdam and New York: Rodopi, 2007. 385 pp. ISBN-13: 978-90-420-2195-2. Reviewed by **Ilka Mindt**, University of Würzburg.

The book reviewed here is one of two publications that depict the wealth of research presented at the 25th ICAME conference, which was held at the University of Verona on 19–23 May 2004. The present book reports on synchronic research into the English language, whereas the other publication (Facchinetti and Rissanen 2006) focuses on diachronic studies.

This volume, edited by Roberta Facchinetti, gives an excellent outline of the state of the art in English corpus linguistics. It is now twenty-five years since the first ICAME conference took place in 1979 in Bergen, two years after the birth of ICAME in 1977 (Svartvik, this volume, p. 22). After more than twenty-five years of research in corpus linguistics, the present book serves as a well-balanced point of reference which documents what has been achieved within this field so far.

The nineteen different papers in this volume are given under three main headings. The first section, headed "Overviewing twenty-five years of corpus linguistic studies", consists of four papers presenting outlines of general aspects of corpus linguistics. The second, headed "Descriptive studies in English syntax and semantics", focuses on a narrower perspective of corpus linguistics in that it comprises research articles dealing with various descriptions in English corpus linguistics. The third section with the heading "Second Language Acquisition, parallel corpora and specialist corpora" offers a wider perspective in that it demonstrates different fields of research where corpus linguistics provides new insights and serves as a powerful resource in other areas of language studies.

The four papers under the heading "Overviewing 25 years of corpus linguistic studies" give a brief outline of different developments within corpus linguistics. Jan Svartvik's biographical article "Corpus linguistics 25+ years on" sketches the history of corpus linguistics in general and of ICAME in particular. In the next paper, Antoinette Renouf examines the development and the creation of corpora from the 1960s up to now. In "Seeing through multilingual corpora" Stig Johansson presents a model of multilingual corpora. He discusses the possibilities multilingual corpora offer and points out some of the problems involved in the study of multilingual corpora. In the paper on "Corpora and spoken discourse", Anne Wichmann stresses the importance of spoken corpora and the wealth of linguistic information they provide. Her primary concern is the availability of spoken recordings. She discusses different annotations of spoken

recordings but convincingly demonstrates that the original sound files are the best resource for studying spoken language.

The eight papers in the section called "Descriptive studies in English syntax and semantics" deal with various aspects of English corpus linguistics and can be grouped into three areas:

a) Corpus architecture;
b) Corpus exploration and development of theoretical linguistic models;
c) Corpus-based studies.

The paper by Mark Davies entitled "Semantically-based queries with a joint *BNC/WordNet* database" is the only one which deals with corpus architecture as its main topic. Davies combines frequency information from the BNC with semantic information from *WordNet* in an interface which allows users to search for different aspects. He explains the architecture of the relational database which contains all relevant data, the properties of the query system as well as the design of the user-friendly interface. Michael Stubbs' contribution, "An example of frequent English phraseology: Distributions, structures and functions" demonstrates how a corpus can be explored in order to offer new insights for theoretical linguistic models. Stubbs analyses multi-word sequences and shows that descriptions based on empirical evidence from corpus analysis can help to formulate theories of language. The other six papers are examples of corpus-based studies focusing on different linguistic aspects. What all the investigations have in common is that the descriptions are based on a combination of frequency information with a detailed qualitative linguistic analysis of the data. Ylva Berglund and Christopher Williams use BNC Baby to describe "The semantic properties of *going to*", showing that *going to* is used in various genres with different meanings. Claudia Claridge focuses on "The superlative in spoken English" and demonstrates that the superlative is not only found in factual comparisons but is also employed as an evaluative expression in involved and emotive language styles. Her data are taken from the spoken part of the BNC. Solveig Granath uses data from the Brown/LOB family and from selected newspaper collections to investigate the different functions of the word *thus*. Granath also demonstrates that for the investigation of certain linguistic aspects, such as the meaning of *thus*, with reference to word order it is important to analyse corpora or text collections which exceed one million words. Rolf Kreyer's research on "Inversion in modern written English: Syntactic complexity, information status and the creative writer" is based on two genres from the BNC. He argues that it is the text producer – the *creative writer* as he calls them – who decides how a

sentence is to be structured and as a consequence chooses to use an inverted construction or not. All other surrounding linguistic factors such as information status and syntactic complexity are in fact a result of that choice. The paper entitled "The filling in the sandwich: Internal modification of idioms", by David Minugh is based on data from the BNC, on a collection of newspapers as well as on *WebCorp*. Minugh shows that corpus data helps in investigating the (though fairly rare and restrained) internal modification of idioms. Liesbeth De Smedt, Lieselotte Brems, and Kristin Davidse analyse 'type' nouns such as *sort of* or *kind of*. Based on a qualitative analysis of corpus examples from parts of the COBUILD corpus and from COLT, they demonstrate how a classification of 'type' nouns can be developed in terms of a functional framework.

Of the seven papers in the third section, "Second Language Acquisition, parallel corpora and specialist corpora", each addresses another aspect within the study and analysis of languages, using corpus linguistic methods. Nadja Nesselhauf considers some results from her analysis of collocations based on the German subcorpus of ICLE. She retraces "the path from learner corpus analysis to language pedagogy" and gives suggestions how research results from learner corpora should be evaluated. Her discussion focuses on the criterion of 'difficulty' and she stresses that this criterion needs to be refined in order to improve teaching. Ron Cowan and Michael Leeser deal with the structure of corpora in SLA research in order to facilitate research on interlanguages. The authors present several error types based on data taken from a corpus consisting of drafts of written English from L2 learners and use them to discuss theoretical considerations in the acquisition of an L2. Francesca Bianchi and Roberto Pazzaglia investigate student writing of research articles in a foreign language. They adopt "a metacognitive/metalinguistic approach to reading comprehension and genre analysis as a prerequisite to the writing tasks." (p. 264). They compile a corpus consisting of psychological studies taken from international journals, which is then used by students as a reference tool for writing in English. Bianchi and Pazzaglia show that the structure of journal articles as well as extracted collocations and phraseological units are important factors in helping students to write more native-like and idiomatic essays. Makoto Shimizu and Masaki Murata concentrate on "Transitive verb plus reflexive pronoun/personal pronoun patterns in English and Japanese" by using a Japanese-English parallel corpus, which consists of newspaper articles and editorials in both languages with the respective translations. The main finding of their study is that co-occurrence patterns of words play an important role in the investigation of language structure. "The retrieval of false anglicisms in newspaper texts" is a research project conducted by Cristiano Furiassi and Knut Hofland, who describe different retrieval tech-

niques for false anglicisms in Italian. The authors show how computational methods employed in corpus linguistics help find language-specific patterns: among them are the comparison of word lists, or the search for phonotactic information. They conclude that false anglicisms are best retrieved by combining automatic and manual procedures. The paper by Josef Schmied on "Exploiting the *Corpus of East-African English*" demonstrates possible uses and limitations of this corpus by applying different software tools. Kerstin Lindmark, Johan Natt och Dag and Caroline Willners adopt corpus linguistic methods in predicting the content of queries on software requirements which have been sent to companies developing software. These queries on software design are written in English and contain information about or requests for improving a particular piece of software. Their data collection consists of the BNC Sampler, a software documentation manual and a database of almost 2,000 queries. The authors present first results that focus on the extraction of a specific terminology for software requirements and try to structure this vocabulary in accordance with the principles found in *WordNet*.

Corpus linguistics 25 years on presents readers with a wealth of possible applications and uses of English computer corpora in the field of synchronic corpus linguistics. It clearly shows what kinds of research results can be achieved by making use of corpus linguistic methods and it also outlines what insights can be gained from the study of the English language through the help of corpora. This volume also opens up many issues and research questions for the coming years. One issue is the difference between a collection of texts and a corpus. Another issue is that the implications of research results for the theory of the English language should be described in a systematic way. A third issue at stake addresses future developments in the use of computational techniques: for example, a wider application of statistical methods, such as the chi-square test, log-likelihood tests, cluster analysis or factor analysis.

All in all, this volume documents in an excellent way what has been achieved within twenty-five years of corpus linguistic research.

References

Facchinetti, Roberta and Matti Rissanen (eds.). 2006. *Corpus-based studies of diachronic English.* Bern: Peter Lang.

Christiane Fellbaum (ed.). *Idioms and collocations*. London: Continuum, 2007. 219 pp. ISBN: 978-0-8264-8994-4. Reviewed by **David Oakey**, University of Birmingham.

The terminology associated with different theoretical approaches to word combinations is itself becoming an area of linguistic enquiry. From the perspective of lexicography, which seeks to order meaning in dictionaries, meaning becomes static in particular combinations referred to as *frozen metaphors*, *frozen phrases* or *fossilized forms*. Cognitive approaches, which are concerned with language processing, production and reception, instead can employ a building metaphor to hint at how such combinations might be stored and retrieved, as in *preassembled speech*, *pre-formulated units* or *ready-made expressions*, although here terms like *syntactic freezes* can also be found. Sociolinguistic perspectives, which highlight the role of word combinations in language use, duly focus on the repetitive, routine nature of the social situations in which they occur, as in *formulaic speech* and *conventionalized forms* (Wray 2002: 9). Computational linguists prefer to work with more literal terms, such as *multiword expressions*, to refer to those combinations whose semantic idiosyncrasies flummox their algorithms.

The extent to which the above perspectives overlap is unclear, and it is debatable whether a common, shared nomenclature is possible, or indeed desirable. At the outset it is therefore worth pointing out that the approach to idioms and collocations in the papers in this book, as made clear by Christiane Fellbaum in her introduction, is closest to that of phraseology and lexicography. The methodological approach does not seek to determine, by applying structural and syntactic criteria, what makes a particular combination an idiom. Instead it describes the syntactic and semantic variation of combinations which have already been identified as idioms. The terminology used by the various contributors reflects their different priorities, and so alongside *idioms* and *collocations*, there are also chapters on *idiomatic multiword units* and *frozen expressions.*

The common source of data used in the work reported in the different chapters in this book is the *Digitales Wörterbuch des deutschen Sprache des 20./21. Jahrhunderts* (DWDS), a reference corpus of the German language constructed at the Berlin Brandenburg Academy of Sciences. These papers deal both with the technical aspects of the corpus, such as its design principles and the methodology by which it was exploited, and the findings of linguistic research into idioms and collocations which it made possible.

The chapter by Geyken begins the collection by outlining the rationale for the DWDS corpus, its design and text selection criteria, its structural and linguistic annotation, and the search engine through which the greater part of the corpus can be publicly accessed.[1] While this chapter is necessarily brief, it still conveys an idea of the impressive amount of work involved in constructing a corpus of this size. The core version of the corpus contains 100 million tokens, and there is a supplementary corpus of around 900 million tokens. Texts in the core corpus are grouped into five genres (27% newspapers; 26% prose, verse and drama; 22% scientific writing, 20% other non-fiction, and 5% speaking) published between 1900 and 2000, containing between two and three million tokens per genre per decade, while the supplementary corpus is primarily made up of recent newspaper data. Although the principal reason Geyken gives for the choice of five genre categories is that "fewer genre distinctions make the daily corpus work easier" (p. 27), he does not rule out increasing the number of categories in the future. A significant number of the texts in the corpus were chosen for the prestige and importance of their authors in relation to other users of German, a restriction which reflects Dr Johnson's insistence on illustrative examples being collected from "masters of elegance or models of style" (Hanks 2005: 264). It could, however, also be argued that a rigorous application of this criterion risks producing a prescriptive corpus which represents the language as the researcher might like it to be used, rather than as it actually is used by its speakers.

The next two chapters report on attempts to avoid the circularity inherent in a corpus-based study such as this one (Tognini-Bonelli 2001), in which the corpus is searched for examples of idioms and collocations taken from the existing literature in order to learn more about why these items are regarded as idioms and collocations. If one is looking in a corpus for examples of a particular idiom, then one needs to have a search item, although it is difficult to formulate a useful search item without already knowing the form of the idiom. It is well known, moreover, that idioms are notoriously variable in form. The English idiom with the canonical form *wash your dirty linen in public*, as Moon (1996: vi) points out, has no stable components at all, and, when searching in a corpus using purely lexical search terms, it would be difficult not to miss some occurrences of this idiom. Corpus queries therefore need to be both flexible enough to catch all relevant occurrences, and also specific enough not to return too many false positives, i.e. pick up forms which are not examples of the intended idioms or collocations. These two chapters accordingly address these issues, and conclude that intuition must still have a role in the construction of corpus queries.

The chapter by Herold discusses the formation of complex regular expressions based on citation forms from idiom dictionaries. Starting with the canonical form of the idiom *jmd. hat Recht*, i.e. *someone is right*, the following query:

```
(NEAR(Recht with [SUB sin]; &hätte; 10) || "@Hätte #10 Recht
with [SUB sin]") && !"[ART] #0 Recht"
```

returns all sentences containing a singular form of the noun lemma *Recht* and the exact forms *hätte* or *Hätte* with no more than ten intervening tokens. At the same time it ignores sentences containing an occurrence of *Recht* which follows an article, and which would not therefore be part of the target idiom. Different queries are needed to yield sentences containing other inflected forms of *haben* in order to reveal the variation of this idiom from its base form. The resulting example sentences then undergo manual sorting to distinguish literal from idiomatic readings, where possible. The chapter by Geyken and Sokirko tests a semi-automatic shallow parsing methodology for classifying noun-verb collocations as verb-nominalization constructions or function verb constructions. They find that it accurately classifies more than 97 per cent of the occurrences of a given verb providing the noun group functioning as its subject is not too complex.

The chapter by Neumann, Körner and Fellbaum describes the online interface, here termed a 'lexical workbench', through which the results of this work on idioms will be made available to future researchers. The user is able to consult a database item, termed a 'template', for a particular idiom which is linked to the examples of that idiom retrieved from the DWDS corpus. The template, based on a MySQL database entry, can be accessed through any standard web browser, and functions like an online dictionary which is linked to an example corpus of sentences containing particular idioms. Each entry provides the user with information about the morphological and lexical variability of the idiom, the text in which it occurs together with its co-text, the genre to which the text belongs, and so on. In addition to this, each entry has several interconnected data sheets containing information on the observed behaviour of the idiom, its dependency structure, morphosyntactic properties, number or tense restrictions, possible lexical substitutions, and possible syntactic transformations, such as whether it can be used in the passive. This is a powerful tool which is likely to be very useful to researchers and learners of German.

The remaining five chapters present the results of studies performed using the above tools and which address different research questions supported by evidence from the corpus. Each chapter focuses on the variation of a different feature of idioms. Stathi's paper focuses on the variations due to adjectival modification: some adjectives modify the main noun in the idiom, while others

modify the whole verb phrase, and still others modify a mixture of the two. Stathi goes on to describe the interplay between the meaning of the adjective and the meaning of the noun or the idiom as a whole. The chapter by Gehweiler, Höser and Kramer focuses on diachronic variation, taking advantage of the century of data in the corpus to investigate how verb-noun idioms have changed in meaning over the years. Their study suggests several reasons why meaning changes: creative uses become current, original meanings are forgotten, or idioms originating in a specialised area of use become used in other contexts. Hümmer's chapter investigates variation in the contextual behaviour of idioms, revealing how this behaviour is influenced not only by semantic factors, such as the literal, metaphorical, and idiomatic meaning of an idiom's component words, but also by formal properties such as the idiom's phrase structure. The chapter by Firenze studies variation in determiners in idioms, and finds them to be less "frozen" than commonly supposed. A determiner can be disagglutinated from its contracted preposition (i.e. *in das* rather than *ins*), deleted altogether, or replaced with an indefinite article or possessive adjective.

Storrer's paper uses the DWDS to test assumptions about a type of verb-noun collocation termed a 'support verb construction', such as *Hilfe leisten* ('to provide help'), in which the support verb is to some extent delexicalised. These forms are traditionally assumed to be interchangeable with base verbs, in this case by *helfen* ('to help'), but Storrer finds that there are in fact restrictions on such interchangeability. Of all these five papers, Storrer's provides the most detailed empirical evidence, in the form of frequency tables as well as examples, to support her case. Finally, Fellbaum gives an account of how the syntactic behaviour of verbs in idioms differs from the behaviour of these verbs when they are used literally. Fellbaum draws an interesting parallel with the form-meaning relationships of verb-noun idioms and those of the grammatical constructions discussed by Fillmore *et al.* (1988) and Goldberg (1995).

The DWDS will be an important resource for corpus linguists, and the idioms 'workbench' tool is likely to appeal to researchers from many different perspectives on word combinations. The papers in this book will be of interest to researchers of German and, since all examples are glossed in English, other languages as well.

Note

1. www.dwds.de

References

Fillmore, Charles J., Paul Kay and Mary C. O'Connor. 1988. Regularity and idiomaticity in grammatical constructions. *Language* 64 (3): 501–538.

Goldberg, Adele. 1995. *Constructions: A construction grammar approach to argument structure.* Chicago: University of Chicago Press.

Hanks, Patrick. 2005. Johnson and modern lexicography. *International Journal of Lexicography* 18 (2): 243–267.

Moon, Rosamund. 1996. Introduction. In R. Moon (ed.). *Collins COBUILD dictionary of idioms.* London: HarperCollins.

Tognini-Bonelli, Elena. 2001. *Corpus linguistics at work*. Amsterdam: John Benjamins.

Wray, Alison. 2002. *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Eileen Fitzpatrick (ed.). *Corpus linguistics beyond the word*. *Corpus research from phrase to discourse* (Language and Computers 60). Amsterdam and New York: Rodopi, 2007. 277 pp. ISBN: 978-90-420-2135-8. Reviewed by **Paul Baker**, University of Lancaster.

Corpus linguistics beyond the word contains a selection of papers from the Fifth North American Symposium at Montclair New Jersey in 2004. Due to space limitations it is not possible to give a detailed description of each paper, but this review instead attempts to summarize some of the main themes that occur across the book. The fifteen chapters are divided into two main sections. Section 1 focuses on analysis tools and corpus annotation, while section 2 is concerned with applications of corpus linguistics – specifically in language teaching and linguistic analysis. The book is generally well-edited and written in an accessible style with simple graphs and tables that are easy to interpret.

Edited collections of papers are often useful indicators of the "state" or progress of a particular academic discipline at a given point in time, and this collection is no exception, demonstrating a maturity in corpus linguistics which is welcome to note. It is heartening to see such a varied collection of papers that use a range of corpus techniques to examine linguistic phenomena above the

lexical level. In the first chapter, as Barrett *et al*. (p. 3) point out, methods that rely on defining text domains based solely on lexical inventory can result in fuzzy boundaries and overlaps. Therefore they hypothesize that certain topics will contain distinct language structures due to stylistic conventions of particular domains, and that a method based on discerning grammatical features (either independently or combined with lexically-based methods) should be considered for domain detection.

It is this theme of going beyond simply lexical analysis which ties together all the papers in the book. So Van Delden discusses the improvement of error rates when using a partial parser and part of speech tagger, Davies details the syntactic annotation of the 1900s portion of the Corpus del Español, and Maynard and Leicher concentrate on pragmatic annotation. Similarly, Vizcaíno also uses pragmatically tagged corpora in a contrastive study of Spanish and English politeness strategies.

It is also good to observe the wide range of statistical processes being carried out on corpus data in the book – particularly impressive here are Barrett *et al.*, who use hierarchical cluster analysis and multidimensional scale analysis in order to distinguish between a variety of text genres, and Deane and Higgins, who employ a singular value decomposition in order to identify latent semantic variables in a text via a vector space model.

Throughout the book a number of tools are discussed, including the Newfoundland part-of-speech tagger (in Barrett *et al.*), SVDPACKC, a piece of software used in dimensionality reduction (in Deane and Higgins), Corpus Coder (in Garretson and O'Connor), Microconcord (in Vizcaíno), COSMASII (in Zinggeler), WordSmith Tools (in Shehzad, and de Haan and van Esch), Biber's tagger (in de Haan and van Esch). There is also some discussion of other tools, created in order to annotate or manipulate corpus data for specific purposes, such as the Java-based tool used by Murzaku and the partial tagger described by van Delden. It would have been useful to have been given more information about the tools used in the latter two papers.

While most of the chapters cover English-based corpora, there are also chapters on Spanish (Davies, Vizcaíno), German (Zinggeler) and Albanian (Murzaku and Jacobson). It is good to see the inclusion of chapters that cover spoken language, which is often neglected by corpus builders, due to issues concerning ethics and transcription. Maynard and Leicher discuss pragmatic annotation of the Michigan Corpus of Academic Spoken English (MICASE), Vizcaíno uses the spoken section of the British National Corpus and the Peninsular Spanish Spoken Corpus while Davis and Russell-Pinson employ the Charlotte Narrative and Conversation Collection.

Two chapters focus on learner corpora – interlanguage analysis (Granger 1998) has become an increasingly popular application of corpus linguistics over the last decade. De Haan and van Esch use a corpus of essays written in English and Spanish by Dutch learners, while Neff *et al*. examine essays written by Spanish learners of English. A related chapter is by Shezad, who uses an EAP (English for Academic Purposes) corpus of computer science articles in order to examine linguistic patterns that are used to outline the structure of an academic paper. Zinggeler's paper on English learners of German does not use a learner corpus but instead focuses on using corpus linguistics techniques in order to create a tool for teaching grammar to learners. Her approach involves getting learners to carry out searches on a corpus of German fairy tales and legends collected by the brothers Grimm. Zinggeler argues that, compared to drill exercises or learning grammar via tables, engaging with an interesting language corpus will result in a more memorable and enjoyable learning experience for students. Echoing the pioneering work on data-driven learning by Tim Johns, Zinggeler notes that students love carrying out detective work as they become language researchers.

Another interesting paper is by Johansson who analyses the use of relativizers (*wh*-forms and *that*) in trials, drama and letters taken from the one million word Corpus of Nineteenth-Century English (CONCE). Relativizer usage is examined in relationship to speaker role and gender, with Johansson finding that the more formal *wh*-forms tend to be favoured in the nineteenth century, although some female letter writers sometimes used the informal *that*, whereas playwrights used *that* in order to index the speech of waiters, maids and servants. The paper makes a good contribution to work surrounding the ongoing informalisation of English, which has mainly focused on the twentieth century (see for example, Leech's (2002) work on the decline of modal usage in American and British English).

In conclusion, this is a well-chosen collection of papers, demonstrating the potential of corpus linguistics for contributing towards phrasal and discourse analysis. It is hoped that *Corpus linguistics beyond the word* will inspire more corpus-based researchers to think beyond the lexicon.

References

Granger, Sylviane (ed.). 1998. *Learner English on computer.* London: Longman.

Johns, Tim. 1997. Contexts: The background, development and trialling of a concordance-based CALL program. In A. Wichmann, S. Fligelstone, T.

McEnery and G. Knowles (eds.). *Teaching and language corpora,* 100–115. London: Longman.

Leech, Geoffrey. 2002. Recent grammatical change in English: Data, description, theory. In K. Aijmer and B. Altenberg (eds.). *Advances in corpus linguistics,* 61–81. Amsterdam and New York: Rodopi.

Marianne Hundt. *English mediopassive constructions. A cognitive, corpus-based study of their origin, spread, and current status* (Language and Computers 58). Amsterdam and New York: Rodopi, 2007. xv + 222 pp. ISBN 90-420-2127-6. Reviewed by **Jürgen Esser**, Bonn University.

The present book is the edited version of a Freiburg habilitation dissertation from 2002. It is a corpus-based study on an intriguing, spreading construction, e.g. *the book sells well*, which is called mediopassive or middle. Within a loose cognitive theoretical framework, the author wants to set her study off from "purely intuition-based" approaches (p. 5). The chapter-structure of the book is as follows: 1 Introduction, 2 Defining the object of study, 3 Previous studies, 4 Theoretical background, 5 The mediopassive in Present-Day English, 6 The history of mediopassives, 7 Conclusions.

It is difficult to introduce the topic of mediopassive constructions since the phenomenon is related to linguistic categories that we traditionally like to separate: clause structure (SVO), semantic roles (agent, patient, animacy), lexical item (polysemy), construction (restrictions of tense, adverbial and general reference), overt and underlying structure. This is how the author introduces her object of study:

> At the core of category, we find intransitive uses of inherently transitive verbs, i.e. verbs where the transitive pattern is the primary one. This does not entail that the verbs themselves are transitive but that they are predominantly used in transitive clauses. Transitivity is taken to be a property of clauses, not of individual verbs. Typically, the object of the transitive pattern occurs in subject position in the mediopassive construction. (p. 7)

Without a sufficient theoretical background Hundt's introductory (not final) definition is not very illuminating. It is confusing to note on the one hand that there are "inherently transitive verbs" (in the footnote called "underlying transitive verbs") and on the other hand that "transitivity is a property of clauses, not of individual verbs" (p. 7). This puzzle and other puzzles are resolved later in Chapter 4.

The comparison of mediopassive and ergative constructions offers many interesting points, e.g.:

> Unlike mediopassives, ergatives can be used intransitively without modification [by manner adverb, JE] and in non-generic contexts. They also usually do not imply an external agent. (p. 11)

This is corroborated by many examples. But again, the theoretical framework is blurred with notions like these "the inherently ergative verb *freeze*". One has to ask: is 'ergative' a property of a construction or of a verb? Therefore, in all, the title of Chapter 2, "Defining the object of study", is rather a misnomer. It presupposes many notions that are explained in the literature or later in the book. A tabular juxtaposition of the features of ergative and mediopassive constructions would have been helpful for the reader.

Also in Chapter 3 Hundt often establishes no common ground between her wide knowledge and that of the reader. The classifications are often implicit and difficult to follow, for example:

> What these [earlier, JE] studies have in common is that they often fail to distinguish between verbs like *open* on the one hand and intransitive constructions of verbs like *sell* and *clean* on the other. (p. 25)

Here, the reader must guess what the properties of *open* are. Quite often, more explanations are wanted, for example when she summarizes Lemmens' (1998) distinction between a transitive and an ergative prototype:

> Instantiations of the transitive prototype include both prototypically transitive processes (e.g. *John hit Mary*) and prototypically intransitive processes (e.g. *Mary is running*). (p. 50)

The author shows that almost all structural, semantic and collocational (frequential) aspects have been described in the traditional structuralist and generative literature. As always, the generative approaches are chiefly concerned with tinkering about with alternative theoretical models and less so with a comprehensive

explanation of the empirical facts. Nevertheless it is noteworthy that the idea of a prototype solution is also expressed in the generative framework.

Hundt favours an approach to the mediopassive construction in the framework of cognitive grammar. Several aspects of mediopassives have already been described in this framework, especially the relation to transitive and intransitive patterns. But Hundt sets out to develop a more comprehensive and unified model.

One can start reading the book with Chapter 4, "Theoretical background", without much loss of information. The prototype approach adopted in this study does not only refer to the construction itself but also to its characteristic properties, namely transitivity, semantic roles and voice. Following Barlow and Kemmer (1994) and Goldberg (1995) the author distinguishes between the semantic transitivity of a verb, i.e. participant roles that are part of an event schema, and syntactic transitivity, i.e. argument roles that are overtly expressed in a clause schema. With these variables the following expressions can be distinguished:

(1) he was sleeping

(2) the glass broke

(3) the book sells well

The prototypical event schema associated with the process of sleeping requires only one participant, that of breaking two and that of selling three participants. Event schemata are related in language-specific ways to clause schemata. Example (1) with only one participant role is a prototypical instantiation of an intransitive clause schema requiring only one argument role. By contrast, the mediopassive construction in (3) is an non-prototypical instantiation of an intransitive clause schema because there are three participant roles of which only one is overtly expressed by an argument role.

Example (3) can be used to show further prototypical, but not necessary properties of mediopassive constructions: (i) They have affected patient subjects. (ii) The verbal action refers to hypothetical or potential processes, in contrast to the prototypically transitive clause, which is realis. (iii) The patient-subject exerts a responsible or controlling participant role. In the words of Hundt:

> The movement of the patient into subject position in mediopassive constructions results in the transfer of agent-characteristics like control and responsibility onto the patient. (p. 68)

And lastly, (iv) mediopassive constructions share with *get*-passives the function that the original agent is taken out of focus.

The author assumes that prototypical mediopassive constructions have a specific pattern meaning which accounts for the productivity of the pattern. The relations to other constructions (transitive, *be*-passive, *get*-passive, reflexive, intransitive construction and ergative) are shown in a network of inheritance links (Figure 4.3, p. 75). Here again, the reader would have benefited from more explanations and illustrative examples.

Chapter 5, "The mediopassive in Present-Day English", describes the corpus linguistic findings of the present study. Starting from thirty verbs that are attested in the literature, the frequencies of these verbs were established in the four standard corpora LOB, FLOB, Brown and Frown. Only five of the thirty occurred with a sufficiently high frequency: *establish*, *read*, *reduce*, *sell* and *wear*. Further empirical sources were American mail-order catalogues, a private collection of example sentences from various sources, and the BNC for specialized searches.

Apparently Hundt works with lemmatized word-forms, that is, for example, *read* in her statistics would also cover word-forms like *reads* and *reading*. She distinguishes between four "transitivity profiles": transitive, intransitive, absolutive and reflexive patterns/uses (p. 88ff.). Apparently the designation 'absolutive' is given to cases of an ellipted object NP (e.g. *John was reading*).

The transitive use of the five verbs is by far the most frequent. Depending on the verb and the corpus, this use is attested in roughly 80 to 95 per cent of the cases, the intransitive use is roughly between one and eight per cent, and the other uses are correspondingly very infrequent. The intransitive cases are the candidates for mediopassive constructions, which instantiate the mediopassive pattern meaning more or less prototypically according to the features (i) to (iv) above.

It should be noted that the statistics about mediopassive constructions depend on the judgement of the analyst. One of the cases that I found difficult to understand concerns the 'bare mediopassive construction'. A clear case is example (4); (5) is excluded because of the following modification. It is, however, unclear to me why (6) should qualify as a bare mediopassive construction although it is modified, cf. pp. 111–113:

(4) Currently black leather jeans and men's frilly shirts are ***selling***.

(5) Brogan shoes ***retailed*** *at prices from $1.25 to $2.25*.

(6) The chair back ***adjusts*** *to several reclining positions*.

Without going into the many interesting statistical details that Hundt has uncovered, perhaps her most important findings are the following:

> The data confirm the hypothesis that prototypical mediopassive constructions are derived from inherently transitive verbs. [...] Intransitive and mediopassive constructions of verbs such as *read*, *sell* and *wear* are clearly derived patterns. [...] The mediopassive is used with a much higher frequency in the language of advertising where it is almost the unmarked pattern for a number of verbs (e.g. *adjust* and *fold*). (p. 126)

The present writer fully supports the cognitive, corpus-based approach adopted by Hundt, but it would have been an improvement to read something about the theoretical status of 'construction', 'inherently', 'underlying', 'derived', 'marked/unmarked' and 'pattern' and how these highly polysemous notions are conceptualised in the cognitive framework. We also miss definitions of 'word-form', 'lemma', 'lexeme' or 'lexical item'. The author offers useful building blocks but there is no integrated model of how contextual features, syntactic structure, lexical item (verb), lexical meaning (homonymy, polysemy) and frequency in the corpus interact.

The terminological imprecision that showed in the preceding chapters cannot be attested in Chapter 6, "The history of mediopassives". Hundt discusses various theories on the origin of the mediopassive and its relation to other constructions, for example, the passival (*the house is building*), reflexive constructions and adjectives in *-able*.

The corpus material consists in the case of mediopassive constructions of four mail-order catalogues from the years 1897, 1927, 1957 and 1986, which were manually searched. Additionally, machine-readable historical corpora (Helsinki, Lampeter, ARCHER and EModE tracts) were used for analyses of reflexive constructions. Furthermore, examples from the linguistic literature and the *OED* were used.

Hundt argues convincingly, mainly on semantic and statistical grounds, that the passival and the mediopassive are not genetically related. Her analysis also suggests:

> [...] that the importance of the discourse frequency of reflexive pronouns for the development of mediopassive constructions has been overrated. (p. 156)

In other words, there seems to be "very little evidence of a systematic variation between mediopassive constructions and reflexive variants." (p. 153) On the

basis of the mail-order catalogues Hundt also rejects the hypothesis of a possible diachronic variation between mediopassive constructions and adjectives in *–able*.

As for the productiveness of the mediopassive construction, the author shows impressively how this phenomenon has increased over the last century in the catalogues. That is, the mediopassive construction is very productive in modern advertising, where inherent properties of the goods are explained.

In all, the book is a diligent study which offers many new insights and a wealth of examples which demonstrate the gradient character of grammatical categories. Hundt shows that, even at our advanced stage of computerization, semantic and pragmatic studies in corpus linguistics still need an attentive linguist at the (wo)man-machine interface.

References

Barlow, Michael and Suzanne Kemmer. 1994. A schema-based approach to grammatical description. In S. Lima, R. Corrigan and G. Iverson (eds.). *The reality of linguistic rules*, 19–42. Amsterdam and Philadelphia: Benjamins.

Goldberg, Adele. 1995. *Constructions. A construction grammar approach to argument structure*. Chicago: University of Chicago Press.

Lemmens, Maarten. 1998. *Lexical perspectives on transitivity and ergativity. Causative constructions in English*. Amsterdam and Philadelphia: Benjamins.

Marianne Hundt, Nadja Nesselhauf and **Carolin Biewer** (eds.). *Corpus linguistics and the web*. Amsterdam and New York: Rodopi, 2007. 305 pp. ISBN 90-420-2128-4. Reviewed by **Gunnar Bergh**, Mid-Sweden University.

Representing the fourth age of corpus linguistics (cf. Kilgarriff and Tugwell 2002), web linguistics is a discipline which is concerned with empirical research based on different forms of language material collected from the web. As such, it represents a practice which has at its disposal the greatest collection of linguistic data ever compiled, i.e. an unprecedented stock of up-to-date, unfiltered electronic text, freely available and maximally broad in topicality, diversity and

domain coverage. Accessed by means of available mining agents, it forms a virtually inexhaustible resource for further advancement in the field of corpus linguistics, presenting as it does an avenue to some of the most central questions about the nature of language use today. Yet, it is also clear that the linguistic usefulness of the web is restrained by its anarchic character, and by the fact that it is constantly changing and growing. Its accidental composition of texts and text fragments thus tends to have a thwarting effect on the systematic exploitation of online data, requiring judicious selection of language material for each individual research initiative (Bergh and Zanchetta forthcoming).

Given the above challenge for corpus linguistics in the twenty-first century – a situation which certainly has attracted both excitement and reluctance among scholars in the field – it is not surprising to find that there is an increasing flow of publications in the literature relating to various aspects of using the web as a source of linguistic data. One such publication is *Corpus linguistics and the web*, edited by Marianne Hundt, Nadja Nesselhauf and Carolin Biewer. This 305-page volume is based on a selection of papers presented at a symposium on "Future perspectives of corpus linguistics" organized in Heidelberg in 2004, and has later been complemented by further papers solicited from leading scholars with an interest in corpus-based research. Altogether, the book contains 15 separate papers divided into four sections: "Accessing the web as corpus" (pp. 7–68), "Compiling corpora from the internet" (pp. 69–132), "Critical voices" (133–166) and "Language variation and change" (pp. 167–305).

In their introduction to the volume, the editors capitalize on the dramatic development of corpus linguistics in recent years. In view of the concomitant increase in standard corpus size from one million words to 100 million words, they bring up the crucial question if this size, embodied for example by the BNC, is sufficient for the wide variety of empirical research ideas today, or if the new multi-billion-word horizons of the web have enough linguistic potential for scholars to sacrifice some of the control and representativeness of traditional sources. The answer given is clearly in the affirmative, although a combination of the two approaches is recommended whenever possible. Among the arguments provided in favour of using the web in this context, they mention the following: (i) a greater diversity with regard to regional variation, (ii) a better coverage of new text types, not least those involving e-mails, chat rooms and blogs, (iii) a greater ease of access to machine-readable text in general, and (iv) a more direct channel to ongoing language change through the existence of "weblish". Furthermore, following de Schryver (2002), the editors make a distinction between two ways in which the web is typically used in current corpus linguistics, viz. either as a corpus itself (Web as Corpus, WaC), or as a source for

compiling a corpus (Web for Corpus, WfC). In their description of these two approaches, they also discuss a number of methodological difficulties caused by the haphazard character of the web, as well as some search problems encountered when using general-purpose search engines in linguistic field work.

The first section of the book, "Accessing the web as corpus", accommodates three papers which are all concerned with practical aspects of performing web linguistics. In their opening paper, Anke Lüdeling, Stefan Evert and Marco Baroni give a survey of the current state of the art of using web data for linguistic purposes. They begin by discussing the pros and cons of various approaches, paying particular attention to the problems of using a commercial search engine as the agent for mining data, and then turn to some of the core issues within this discipline, notably such pertaining to the quality and nature of collected online material in terms of quantity, representativity, variation and stability, among other things. A considerable part of the discussion is devoted to the desideratum of a powerful linguist's search engine, i.e. a web agent that would be able to crawl, post-process, annotate and index a sizeable portion of the web, thereby aiming to provide corpus linguists with better control of collected data.

Continuing the descriptive efforts, William Fletcher delves deeper into the characteristics of the web by considering its size, composition and evolution, as well as its rewards and limitations as a linguistic corpus (WaC) and as a source for one (WfC). He elaborates on the methodological aspects of concordancing the web – its "promise and problems, tools and techniques" (p. 25) – making particular reference to the development and (dys)functionality of general search engines. As an alternative to these common gateway applications, he introduces his own concordancing software, KWiCFinder, described as an easy-to-use hunting and grazing agent which is able to conduct web searches, retrieve matching documents, and produce interactive concordances of search terms. While still noting the many pitfalls of using such "webidence", he concludes that, as methods improve to ensure the quality of data, the web has good prospects of eventually becoming a fully legitimate source for corpus linguistic research.

Another practical contribution to the methodology of web linguistics is described by Antoinette Renouf, Andrew Kehoe and Jayeeta Banerjee. Recognizing the potential of the web as an inexhaustible source of up-to-date text in various languages, they report on the WebCorp initiative, a long-term development project which has produced a corpus tool able to extract linguistic data from web text, and to present them in a way similar to that of finite corpora. Yet, as the authors acknowledge and discuss, there remain many linguistic and procedural problems with this type of application, caused either by the lack of stan-

dards on the web, or by the constraints and time lags imposed by mediating search engines. As a way to improve performance, the project team has now initiated work to develop a linguistically tailored search engine in which WebCorp will play an even more central role.

The second section of the book, "Compiling corpora from the internet", is concerned with WfC aspects, and contains three papers on the construction and use of specialized corpora compiled from different types of online material. Focusing on the genre of news programme language, Sebastian Hoffmann discusses technical and methodological issues of creating a large corpus of spoken data based on public CNN transcripts collected from the web. He gives an outline of the different steps necessary in converting the contents of such downloaded documents into a format compatible with standard concordancing software, illustrating the potential use of the outcome through a sample analysis of the intensifier *so* in different syntactic environments. The results of his investigation show that, although there are many difficulties with WfC data mining, such customized corpora can indeed be used profitably to complement available corpora in studies of present-day English.

Claudia Claridge pursues a similar perspective by reporting on her project of creating a special corpus of message board (forum) language collected from "electronic agora" on the web, in particular one that takes regional variation into account. Arguing that this type of language makes up a fast-growing genre of its own at the crossroads of speech and writing, she describes the technical aspects of transforming such postings into an annotated corpus, and then goes on to discuss material-specific problems, for example how to account for the sequencing of online messages and for the regional identity of speakers involved. In a pilot study of a set of interaction and attitude markers, among them personal pronouns and related speech act items, which are seen as central features of this text type, she shows the potential of compiled forum language as a means to go beyond traditional corpora in studying recent variety-specific usages in English.

The section is concluded by a multi-dimensional analysis of online text categories carried out by Douglas Biber and Jerry Kurjian. Sparked by the observation that the usefulness of the web as a linguistic source is often limited by difficulties in identifying the text category of downloaded documents, they collected a stratified 3.7-million-word corpus from the web categories "Home" and "Science" to compare the power of two analytical approaches to the problem, one which is based on Google's predefined categories, and another which uses text types proper as defined through selected linguistic criteria. The dimension scores of the study show that the two text categories at hand are not clearly distinguished on linguistic grounds, which is tantamount to saying that they are less

useful for corpus linguists. Rather, the scores suggest that the compiled material contains no less than eight different text types, identified and interpreted in terms of their salient linguistic and functional characteristics.

Bearing the title "Critical voices", the third section of the book is couched in a more evaluative vein, featuring two critical appraisals of the current state of corpus linguistics and its inherent possibilities. Geoffrey Leech brings up the relationship between old and new language resources in the light of traditional corpus requirements such as representativeness, balance and comparability, suggesting that many empirical linguists of today seem to have sacrificed these yardstick notions on the altar of practicality, pragmatism and opportunity. The easy and fast access of web-based material is thus said to have had a tendency to limit and skew our research efforts to language data which are readily available by mouse click rather than being theoretically interesting in themselves. While still embracing new developments in this field, recognizing both potential and limitation of the web as an added resource, he stresses the need for us to improve and refine data collections and methods that we already possess, all in order to accentuate the importance of carrying out research on corpora compiled according to design and systematic sampling.

Graeme Kennedy follows suit by claiming that "bigger is not necessarily better" in corpus linguistics (p. 152), and that the web has yet to prove its advantages over large corpora that have been carefully constructed. True to his stance, he brings in the BNC as a case in point, arguing that the richness of this corpus is partially under-exploited for the description of English and for the related processes of language learning/teaching. He bolsters his case by a statistical study of the semantic relations in collocations containing the verbs *find* and *lose* as well as selected amplifiers, showing that such structured data have a bearing not only on the explicit knowledge of language learners, but also on the implicit curriculum that language imposes on them. It is through such exploitation of balanced corpora, the author claims, that we may eventually find a means to tackle the web as a source for building huge monitor corpora.

The final section of the book, "Language variation and change", is also its most comprehensive part, containing seven case studies on such different topics as morphology, syntax and lexis, as well as synchronic and diachronic variation in English. Evidencing both WaC and WfC approaches, these studies typically show that the mass of textual data from the web can provide crucial evidence in many research questions, not least if the results are combined with those from standard corpora.

Anette Rosenbach reports on her study of grammatical variation in present-day English, specifically the interplay between certain *s*-genitives and noun-

noun constructions, e.g. *driver's licence* vs. *driver licence*. As it is difficult to find a sufficient number of relevant data in traditional corpora, she turns her attention to the web through the Google and WebCorp interfaces, showing that there is a clear gradience between the two target variants in the sense that the animacy of the modifier typically determines the choice of construction. More importantly, however, her study highlights the general benefits and problems of mining grammatical data on the web, and demonstrates the specific advantages that a linguistically tailored system such as WebCorp brings in this context.

Günter Rohdenburg sets as his main task to compare the output of two different resources in corpus linguistics, the web data provided by Google and the large newspaper corpora available at Paderborn. The framework of his study consists of four variation principles in English, tested heuristically through the Google agent, namely that (i) explicit options are preferred in cognitively more complex environments, (ii) unmarked infinitives are less prone to allow extraction than marked infinitives, (iii) juxtaposition of formally identical or near-identical grammatical structures is avoided, and (iv) variants such as *scarved* and *leaved* are more strongly attracted to plural contexts than their rivals *scarfed* and *leafed*. On all four counts, the results confirm the predicted tendencies, indicating that the distributional patterns are determined by functionally motivated, and presumably universal, tendencies. It is also shown that the two sources of data are strikingly parallel in their achievement, making a case for the usefulness of web linguistics in this field, despite the relative "messiness" of data and the lack of sophisticated search tools.

Britta Mondorf takes up the cudgel for the web as a means to study semantic, pragmatic and cognitive factors that are recalcitrant to empirical testing even in conventional mega-corpora. Her main vehicle for doing this is the comparative construction in English, in particular the well-known competition between synthetic and analytic forms, as in *friendlier* vs. *more friendly.* Using a combination of a 600-million-word collection of corpora and the multitude of textual data available on the web, she offers support for the idea that a theory of processing efficiency can best explain the morpho-syntactic variation involved in this context, not least because abstract concepts, which are thought to involve a higher processing load, tend to favour the analytic construction. She also makes reference to the possibilities of using web data as a source for historical analyses, specifically in connection with cases of iconic ordering of coordinated comparatives. Echoing Rohdenburg's conclusion, she notes that there is considerable overlap in the patterns derived from corpora and web data, a finding which suggests that accessing the web provides promising avenues for future linguistic research.

Assuming a more general perspective on empirical research methodology, Christian Mair discusses the increasing importance of the web as a source of data for linguistic studies of ongoing change and recent usage. Since closed corpora often paint an incomplete or distorted picture of the current situation, he argues in favour of the web – "the accidental corpus" (p. 236) – as the natural remedy to such problems, with its virtually unbounded amount of up-to-date textual material from different registers. Concentrating on prepositional usage with the adjective *different* as well as the distribution of the past perfect passive progressive and the *save (from) V-ing* construction in English, he demonstrates, to different extents, that regional variation data from closed corpora can be replicated through domain-specific searches of the web, e.g. such restricted to .uk, .us, .edu or .gov. Hence, while the odds may seem tremendous, his conclusion is still that the rough-and-ready procedures of web-based research can be successful, expressly in the case of variation issues of "low and medium levels of delicacy" (p. 244).

Marianne Hundt and Carolin Biewer, two of the editors of the volume, expand the discussion of regional variation by bringing up the possibilities of using the web when studying varieties in the South Pacific and East Asia. In particular, they set out to investigate whether the development of the (inner circle) varieties of English in Australia and New Zealand shows any noticeable modelling effect on their neighbouring (outer circle) varieties of English, such as those spoken in the Philippines, Singapore and Fiji. To this end, they applied a WfC approach to collect a large number of articles from online newspapers in the latter varieties, thereby forming the so-called South Pacific and East Asian Corpus (SPEAC). With variation between the past tense and the present perfect as the dependent measure of their case study, they found, contrary to expectations, that there was no evidence in favour of a growing influence of the inner circle varieties on the collected corpus material. However, this finding does not necessarily disqualify the methodology as such, but rather suggests that a more thorough lexico-grammatical basis is needed for future study in the field.

Another attempt to employ web data for research on non-standard English is reported by Lieselotte Anderwald. Her investigation concerns non-standard past tense verbs, such as the imperfect forms *rung* and *drunk* (referred to as Bybee verbs, from Bybee 1985), and their possible usage in present-day informal English. Drawing her basic distributional data from the Freiburg English Dialect Corpus (FRED), she relates those primary figures to search results mined from the web domain .uk, first through WebCorp and then (somewhat more successfully) through Google. The results show that these typically dialectal forms are still in frequent use in current non-standard English, a phenomenon which is

said to be attributable to both historical continuity and the principle of functional analogy.

The final contribution to this section comes from Nadja Nesselhauf, the third editor of the volume. With the aim of exploring the possibilities of using web data also for diachronic analysis, she applies a research paradigm involving the future time expressions *will*, *shall* and *'ll* in selected linguistic contexts, which are studied first in the nineteenth-century British English material of the ARCHER corpus, and then in a collection of contemporary fiction texts downloaded from the web, referred to as WebFict. The most noticeable result of the investigation concerns the development of the contracted form *'ll*, which exhibits a decrease in frequency in the nineteenth century as measured by the ARCHER corpus, but an increase as measured by WebFict. This discrepancy is likely to be partially due to intertextual variation in the use of this form, but may also highlight the methodological problems of such comparisons. Yet, the bottom line of the study is that "a quick-and-dirty corpus from the web" (p. 287) can yield good insights also into the realm of diachronic linguistics, although it is advisable to supplement such findings with data from a traditional corpus.

Turning now to some evaluative comment, it is clear that the volume *Corpus linguistics and the web* makes up a valuable contribution to corpus linguistics in the fourth age. With its general approach to both potentials and problems in web linguistics, it fills an important gap in the description of an auspicious research methodology which is zooming rapidly into the twenty-first century with a fair share of growing pains. One virtue of the book, for example, is its balance of contents, viz. the fact that it captures the good prospects of the web as a source for linguistic research, while still keeping a critical perspective on its range of usefulness, thus avoiding the fallacy of undue praise at a novel methodology (or "the-emperor's-new-clothes effect"). Another virtue has to do with its joint treatment of the WaC and WfC approaches, the two main applications of web data, which provides some good spot tests of the different possibilities of using the web as a massive but undressed reference corpus as well as a source for building customized corpora from select online archives. A third virtue is realized through the prudent compromising tone of many of the WaC contributors, emphasizing the need, at least in the present state of the art, to combine results from web-based studies with such from traditional corpora, all to the purpose of creating a more solid empirical basis for making qualitative and quantitative claims of new linguistic discoveries.

However, there is also a down side to the present volume. One such aspect, for example, is the tendency towards overlap between the different papers when describing the web as a linguistics resource and the reasons for using it in this

capacity. Inevitable though it might be in the context of conference papers using similar methodologies, this repetitiveness is slightly disturbing for the reader in the sense that one and the same message tends to be conveyed several times. Another, more important problem concerns the WfC perspective of the book. While there is a relatively extensive coverage of WaC research and related search tools, the treatment of corresponding WfC aspects is somewhat narrow, *pace* the good efforts of the two initial papers, specifically when it comes to the description of more elaborate crawling and post-processing strategies, the use of the web as a test bed for the training of automatic search tools, and the building of disposable parallel corpora in the context of machine translation. In addition, it is clear that the book would have profited from a more consistent organization. One case in point is its somewhat ad hoc division into sections, which is clearly stated in the introductory part but is nowhere else to be seen among the following 300-odd pages, an organization which also subsumes a sometimes artificial clustering of papers, at least as far as the different topics are concerned.

In the critical aggregate, however, there is no doubt that the positive impression of the book prevails. This is so in particular as it offers a wealth of insight into common approaches to web-based language study, with its strength lying in the manifold treatment of web methodology, often in conjunction with traditional corpus methods, and in its variety of interesting research results, either in a WaC or WfC framework. Thus, despite some noted shortcomings, this publication constitutes another important step in the establishment of web linguistics as the currently most rewarding approach in corpus linguistics.

References

Bergh, Gunnar and Eros Zanchetta. Forthcoming. Web linguistics. In A. Lüdeling and M. Kytö (eds.). *Corpus linguistics: An international handbook (Handbücher zur Sprach- und Kommunikationswissenschaft [HSK] / Handbooks of Linguistics and Communication Science*). Berlin: Mouton de Gruyter.

Kilgarriff, Adam and David Tugwell. 2002. Sketching words. In M-H. Corréard (ed.). *Lexicography and natural language processing: A Festschrift in honour of B. T. S. Atkins*, 125–137. Göteborg: EURALEX.

Schryver, Gilles-Maurice de. 2002. Web for/as corpus: A perspective for the African languages. *Nordic Journal of African Studies* 11: 266–282.

Stig Johansson. *Seeing through multilingual corpora: On the use of corpora in contrastive studies* (Studies in Corpus Linguistics 26). Amsterdam and Philadelphia: John Benjamins, 2007. xxi + 355 pp. ISBN 978-90-272-2300-5. Reviewed by **Lars Borin**, Göteborg University.

Seeing through multilingual corpora: On the use of corpora in contrastive studies is a work which summarizes and extends more than a decade of contrastive corpus linguistics research conducted by Stig Johansson and his group at the University of Oslo, in the English-Norwegian Parallel Corpus (ENPC) and Oslo Multilingual Corpus (OMC) projects. The book is divided into fifteen chapters, four of which contain material of a more general theoretical and methodological nature. Here Johansson motivates the use of corpora in contrastive studies (Chapter 1), describes how the three multilingual corpora used for the studies reported in the book (the ESPC, the OMC and a small multiple translation corpus) were constructed (Chapter 2), discusses in more detail what sorts of research questions in contrastive analysis and translation studies can profitably be addressed with the help of multilingual corpora and what kinds of tools are needed for conducting this research (Chapter 3), and, finally, gives a picture of the state of the art of multilingual corpus research and its prospects for the future (Chapter 15).

These four chapters form a kind of theoretical and methodological frame for the remaining eleven chapters (Chapters 4–14) which contain detailed investigations of various contrastive linguistic problems. This division of the book into a theoretical framework part and concrete empirical linguistic investigations coincides with another division of the material in the book into a largely newly written part (Chapters 1–3 and 15) and another part consisting of (revised versions of) previously published texts (Chapters 4–14), thirteen journal articles and book chapters originally published between 1997 and 2006.

The investigations are of three kinds, going from the particular to the general (the following classification is mine, not Johansson's):

(1) Several studies concern the behaviour of (open-class) lexical items in a contrastive perspective. Classified by part of speech in the source language, there are studies of nouns: words for times of the day in English, Norwegian, German and French; English *mind*, *thing* and *fact*; Norwegian *menneske* ('human being, person') (Chapter 4) – and verbs: *hate* and *love* (Chapter 5), *spend* [time] (Chapter 6) and *seem* (Chapter 7). These lexical topics take up about 40 per cent of the part of the book devoted to individual linguistic studies (98 out of 258 pages). Nouns and verbs receive about equal attention, thus coun-

teracting a prevailing inordinate fondness for verbs in contrastive corpus studies (p. 35f.).

(2) About thirty per cent (77 out of 258 pages) is devoted to studies of grammar and discourse as reflected in the behaviour of individual text words and text word combinations (typically closed-class items and function items): what Johansson refers to as 'usuality'[1] (Chapter 8), Norwegian *man* vs. Eng. *one* (generic person pronoun) (Chapter 10), Norwegian *likevel* 'still, anyway, after all' (Chapter 13; co-authored with Thorstein Fretheim) and *well* (Chapter 14).

(3) Finally, in the remaining thirty per cent (83 out of 258 pages), we find studies on grammar and discourse where the point of entry is through a grammatical category or grammatical pattern more directly: (clausal) negation (Chapter 9), subject choice in translations (Chapter 11) and sentence openings in translations (Chapter 12).

All the studies involve comparisons between Norwegian and English. In addition, French, German and Swedish data are used in several studies. In the studies investigating subject choice and sentence openings in translations, a multiple-translation corpus is used, where several professional translators have independently translated the same two English texts into Norwegian. As befits a work in corpus linguistics, all studies are amply illustrated with examples taken from the corpora described in Chapter 2. Many of the studies are explicitly exploratory and the results indicative rather than definitive, pointing out directions for further research.

Thus, concrete research problems in contrastive analysis and translation studies are the focus of Johansson's book, and students of these fields will find a wealth of invaluable data and interesting results in it. Some general conclusions about languages in contrast which receive repeated confirmation throughout the book are the following.

Translation effects are real. Time and again, Johansson finds evidence that translated texts deviate systematically from comparable original texts in the target language, in ways which reflect linguistic characteristics of the source language.

The major genres fiction and non-fiction are clearly distinguished linguistically both in original texts and translations.

Lexical items are much more complex than even the best dictionaries would have us believe. In his book *Language*, Sapir (1921: 39) stated that "all grammars leak". Now corpus studies empirically confirm the long-held belief by some linguists and lexicographers (perhaps most vociferously by the Moscow school of lexical semantics and lexicography, as represented by e.g. Mel'uk and Žolkovskij (1984) or Apresjan (2000)) that all dictionaries deceive. From

Johansson's studies we learn that bilingual dictionaries deceive doubly (Section 15.5.1), with potential consequences for lexicographic practice and translator training (Chapter 15).

Source-language grammatical words tend to show a wider range of correspondences in translation into another language than lexical words, but this is a cline, rather than a clear-cut dichotomy, since the boundary between lexis and grammar is anything but sharp. Thus, I could well have drawn the dividing line differently between the studies in (1) and (2) above, with at least the studies involving *fact*, *thing* and *seem* classified under (2) instead of (1), since these studies actually focus on function-word like uses of these items (e.g., *fact* in the expressions *the fact that* and *in fact*).

The volume is something in between a monograph and a thematic collection of articles. There is a strong thematic strand running through it: the previously published studies have obviously been edited for greater overall coherence, all references have been collected at the end of the book, an author index and a subject index have been added, etc., but the general impression on the reader is still one of a collection of loosely connected articles rather than a unified book-length whole. For example, there are few cross-references between the individual linguistic studies in Chapters 4–14, and it is evident that the corpora and corpus tools used have developed considerably over the ten-year period covered by the studies.

On this note, one could also have wished for a deeper discussion of theoretical and methodological issues to balance out the rich empirical material. While there are such discussions sprinkled throughout the individual studies, in addition to the newly written introductory and concluding chapters devoted to theory and method, the book would have benefitted from an even more unified treatment of these issues, for instance in a more substantial introductory or concluding part.

Although the book raises a number of theoretical and methodological questions all worth a deeper discussion, for reasons of space I will here focus on a set of issues which have direct bearing on contrastive analysis as "the systematic comparison of two or more languages, with the aim of describing their similarities and differences" (p. 1), and where I think there is significant and so far largely untapped potential for synergy between two separate linguistic traditions.

In Borin (2002) (quoted on p. 305), I note the fairly clear division of linguistic work with corpora into two distinct traditions, one using corpora for time-honoured traditional linguistic research and the other using them in developing natural language processing applications (although the former also generates

practical applications – Johansson mentions several in Chapter 15 – and the latter also involves a good deal of quite theoretical work). In the same way, it seems to me that scholars interested in comparing languages also belong to two distinct traditions. One is represented by the work under review; both contrastive analysis and translation studies as defined by Johansson belong here. Work in the other tradition is conducted under labels such as *language typology* (or *linguistic typology*), *areal linguistics*, *contact linguistics* and *grammaticalization*, all of which in practice have turned out to be strongly interconnected areas of research. It is even arguable that contrastive analysis constitutes a special case of language typology and that translation studies make up a sub-branch of contact linguistics. Just like the two corpus linguistics traditions that I discuss in Borin (2002), these two traditions could also interact more, a case in point being that even though there is mention of typology (p. 36) and grammaticalization (Chapter 7) in the work under review, its subject index does in fact not list either.

A contrastive linguistic analysis aspiring to some degree of universality could do worse than look to language typology, contact linguistics and related disciplines for theoretical insights of a more general nature, since these areas have developed at a furious pace in all respects in the half-century that they have existed as modern linguistic disciplines (reckoning from the publication of Weinreich (1953) and Greenberg (1963), respectively). In particular, *lexical typology* is emerging as an increasingly active sub-field of language typology (e.g. Koptjevskaja-Tamm forthcoming), with obvious relevance for several of Johansson's studies, while his studies, on the other hand, can provide detailed data for lexical typological studies, something that there can never be too much of. Language typologists investigate how cross-linguistic variation correlates with particular linguistic phenomena, a paradigm case being pronouns, especially indefinite pronouns, which turn out to show great variation even among closely related languages (Saxena 2006), and thus *a priori* would be expected to show little cross-linguistic correspondence in a study of Norwegian pronouns contrasted with English ones, as in Johansson's study of clausal negation which involves some negative pronouns (Chapter 9).

One issue which bears directly on Johansson's studies, and which has occupied the minds of typologically oriented linguists for a long time, is the teasing out of how general linguistic mechanisms, intrinsic to languages, and language contact interact. To put the matter differently: in a contrastive study using Norwegian texts translated from English, what putative translation effects are translation effects in a narrower sense, and not due to, e.g., genetic closeness, areal effects or language contact in a sociolinguistic setting where a considerable share of all written texts are translations from English and where English is rapidly acquiring the status of a second language rather than a foreign language?

What translation effects will we find in translations from, say, Arabic, Chinese or Malayalam into Norwegian, and why?

Language typologists are showing a fledgling interest in multilingual corpora (e.g. the MPI Leipzig meeting in 2005 on *Parallel Texts: Using Translational Equivalents in Linguistic Typology,*)[2] but so far with a marked lack of input from modern corpus linguistics, which could lead to a great deal of unnecessary reinventing of methodological wheels.

Language typology has sometimes been criticized for relying excessively on second-hand data, in the form of often quite shallow, summary and sometimes unreliable grammatical descriptions of languages. A closer interaction with corpus linguistics could go some ways toward remedying this deficiency. So I would suggest that language typology be added to the disciplines listed as needing closer interaction in the section on future directions (Section 15.6). I predict that this would benefit both fields greatly.

There is a marked difference in emphasis between language typology and corpus linguistics, which would need to be addressed first, however: while students of linguistic typology focus mainly on grammatical phenomena, corpus linguists thus far have tended to work in the area of lexis. Arguably, what makes corpus linguistics something more than – or at least distinct from – other forms of empirical linguistic inquiry, are the computational tools wielded by corpus linguists of all persuasions, which allow them to "organize huge masses of data", giving access to "facts about language use which no amount of introspection or manual analysis could discover" (Stubbs 2002: 221, cited on p. 1). Johansson depicts the use of a corpus in linguistic investigations as "a kind of dialogue between the researcher and his/her material" (p. 38). The tools provide the language in which this dialogue is conducted. Adding a Whorfian twist, we could say that the tools also bear strongly on what kinds of research questions can be asked, and consequently, in Johansson's own words: "lexical patterns are relatively easy to identify in corpora" (p. 35); and: "we can observe the behaviour of word forms and word sequences" (p. 306). The most grammatically oriented of Johansson's studies were conducted either by searching automatically for a small set of text words which were thought to identify interesting grammatical phenomena, often with a good deal of manual post-processing, or by completely manual inspection of (the sentence units in) small text materials.

There is actually a conspicuous lack of studies, for any language, where grammatical phenomena are addressed directly on the large scale that we have come to expect from corpus studies. To see better why this would be so, let us perform the *Gedankenexperiment* of trying to conduct corpus studies on English text where all sentence-internal spaces have been removed. ("Don'tyouknow-

whatthatis? It'sspringfever.Thatiswhatthenameofitis."). Many lexical studies would probably still be feasible, whereas studies of grammar through function words would become very cumbersome, at least using KWIC string searches. But this is the normal situation with many languages, where grammar is encoded as bound morphology, i.e. (often short) pieces of (graphic) words. Usuality (often referred to as the *habitual*), clausal negation, generic or impersonal agents, and evidentiality (one function of *seem*) – i.e. some of the grammatical phenomena investigated by Johansson – are all commonly expressed in this fashion in the languages of the world. Of course, one should not look a gift horse in the mouth; naturally one should take advantage of the corpus-linguistics friendly nature of the linguistic structure and orthography of English and other similar languages, but if we want to extend the use of multilingual corpora in contrastive studies, we will need tools that allow us to access grammatical features of a language directly and wholesale, which will present a real challenge. Johansson is well aware of this (p. 306). He emphasizes *equivalence* and the concomitant *comparability* as crucial issues for future corpus-based cross-linguistic research, especially if we wish to move into the realm of grammar and discourse. Notwithstanding occasional pessimistic statements to the contrary (e.g. Haspelmath 2007), linguistics itself is in fact predicated on the possibility of comparing units and categories across languages, so as linguists we have to believe that this will be possible.

However the future turns out with respect to closer collaboration among some or all of linguistic typologists, contrastive linguists, students of translation and corpus linguists, Stig Johansson has done all these fields a great service in giving us this volume.

Notes

1. 'Usuality' is linguistically marked by the Norwegian expression *det hender*, lit. 'it happens'.
2. http://email.eva.mpg.de/~cysouw/meetings/paralleltexts.html.

References

Apresjan, Jurij. 2000. *Systematic lexicography.* Oxford: Oxford University Press.

Borin, Lars. 2002. ... and never the twain shall meet? In L. Borin (ed.). *Parallel corpora, parallel worlds*, 1–43. Amsterdam and New York: Rodopi.

Greenberg, Joseph (ed.). 1963. *Universals of language*. Cambridge, Mass.: MIT Press.

Haspelmath, Martin. 2007. Pre-established categories don't exist: Consequences for language description and typology. *Linguistic Typology* 11(1): 119–132.

Koptjevskaja-Tamm, Maria. Forthcoming. Approaching lexical typology. In M. Vanhove (ed.). *From polysemy to semantic change: A typology of lexical semantic associations*. Submitted to Benjamins.

Mel'ǔk, Igor and Aleksandr Žolkovskij. 1984. *Tolkovo-kombinatornyj slovar' sovremennogo russkoga jazyka: Opyty semantiko-sintaksiceskogo opisanija russkoj leksiki [Explanatory combinatorial dictionary of modern Russian: Semantico-syntactic studies of Russian vocabulary]*. Wien: Wiener slawistischer Almanach.

Sapir, Edward. 1921. *Language: An introduction to the study of speech.* New York: Harcourt, Brace and Company.

Saxena, Anju. 2006. Pronouns. In Keith Brown (ed.). *Encyclopedia of languages and linguistics*, 2nd edition, 131–133. Oxford: Elsevier.

Stubbs, Michael. 2002. *Words and phrases. Corpus studies of lexical semantics*. Oxford: Blackwell.

Weinreich, Uriel. 1953. *Languages in contact. Findings and problems*. The Hague: Mouton.

Merja Kytö, Mats Rydén and **Erik Smitterberg** (eds.). *Nineteenth-century English: Stability and change.* Cambridge and New York: Cambridge University Press, 2006. xix + 295 pp. ISBN 978-0-521-86106-9, 0-521-86106-3. Reviewed by **Andrea Sand**, University of Trier.

The volume *Nineteenth-century English: Stability and change* is yet another indication of the growing interest in the recent history of the English language, as expressed by publications such as Bauer (1994), Beal (2004) or Mair (2006). Because of its perceived modernity, nineteenth-century English has long been neglected in terms of linguistic inquiry, as the editors of the present volume point out in their "Introduction" (pp. 1–16). Their collection of papers aims at (and succeeds in) filling some of the gaps which still exist despite the publica-

tion of overviews of nineteenth-century English such as Bailey (1996) or Görlach (1999), or in-depth studies of particular grammatical features, such as Smitterberg (2005) on the progressive. Understanding the most recent past helps to explain the present situation, in terms of long-term stability or accelerated change.

All contributions to the volume share a corpus-based methodology. As nineteenth-century corpora are scarce, they mainly rely on the Corpus of Nineteenth-century English (CONCE), a 1-million-word corpus of nineteenth-century English compiled at the Universities of Uppsala and Tampere. The CONCE corpus is subdivided into three periods (1800–1830, 1850–1870 and 1870–1900) and contains several text-types, with drama, debate and trials representing the more oral genres, and fiction, letters, historical and scientific monographs the written genres, both private and public.[1] This allows for text-type specific analysis, albeit based on a relatively narrow database. Because of these restrictions with regard to size and genres, the contributions by Christian Mair and Tony Fairman include additional data, drawn from the *OED* on CD-ROM quotation base (Mair) and from a corpus of mainly letters and bills from English Record Offices (Fairman).

The editors group the contributions to the volume according to their focus on stability or language change with regard to the features under analysis. In my discussion of the individual papers, I will thus follow their grouping based on content instead of the alphabetical order of appearance in the book.

Five of the ten contributions chart language change in progress. The first of these is Ingegerd Bäcklund's study of "Modifiers describing women and men in nineteenth-century English" (pp. 17–55) on the basis of three text-types from CONCE, namely drama, fiction and letters. To determine whether the choice of premodifying adjective phrases, e.g. *good-natured*, or postmodifying *of*-phrases, e.g. *of fortune*, reflect changing gender roles in the nineteenth century, Bäcklund used qualitative analysis, i.e. semantic categorization, and quantitative analysis, i.e. comparison of frequencies, with regard to reference to males or females across the three periods represented in CONCE, in comparison to a study of eighteenth-century usage. Bäcklund's analysis shows that there are indeed differences in the description of men and women, and that there are changes throughout the nineteenth century with regard to the description of women, for example with regard to their intellectual powers. There are also differences in the use of modifiers by male and female writers, reflecting their status and roles in society.

Peter Grund and Terry Walker examine "The subjunctive in adverbial clauses in nineteenth-century English" (pp. 89–109) based on CONCE data, by

comparing clearly identifiable subjunctive forms with indicative and modal constructions across genres, periods, gender of the author and according to specific conjunctions and verbs. They found that the use of the subjunctive decreased considerably throughout the nineteenth century, especially with verbs other than BE and in the less formal genres of trials, letters and drama. The results thus provide "the missing link" between the research on Early Modern English (EME), when subjunctives were still used rather freely, and the rather restricted Present-Day English (PDE) uses of the subjunctive.

Along similar lines, Merja Kytö and Suzanne Romaine analyze "Adjective comparison in nineteenth-century English" (pp. 194–214). Their study reveals a steady increase of the inflected forms (e.g. *happier*) at the expense of periphrastic forms (e.g. *more happy*) in the nineteenth century. The results are less conclusive with regard to register variation due to the internal heterogeneity of the different genres. However, inflectional superlatives are especially common in private letters, due to the stylistic requirements of opening and closing formulae (e.g. *My dearest Mrs. Martin*). With regard to their syntactic position and function, the use of comparative forms comes close to PDE usage by the end of the nineteenth century. As was the case with the subjunctive, the nineteenth century bridges the gap between EME and PDE usage.

The two remaining papers in this first category focusing on language change are concerned with more narrowly defined features of English morphosyntax. Christian Mair looks at "Nonfinite complement clauses in the nineteenth century: The case of *remember*" (pp. 215–228) comparing gerundial and infinitival complements on the basis of the *OED* on CD-ROM quotation database, as both CONCE and ARCHER did not yield enough conclusive evidence for the constructions under analysis. Mair convincingly argues that yet again the foundations of PDE usage were laid in the nineteenth century, when the formerly common constructions with retrospective infinitives were phased out and gerundial constructions gained more ground.

Finally, Juhani Rudanko reports on "The *in –ing* construction in British English, 1800–2000" (pp. 229–241), e.g. *The titans delight in upsetting the odds*, on the basis of CONCE, LOB and two British subcorpora from the Bank of English Corpus. Comparing the matrix verbs used in this construction, Rudanko shows that a shift has taken place from the nineteenth century preference for matrix verbs with a meaning of 'engaging in an activity' to a number of additional semantic groups in the twentieth century. Differences can also be found with regard to the implied subject of the nonfinite clause which was invariably coreferential with the subject of the matrix clause in the nineteenth century data.

Three contributions focus on stability rather than change, among them Larisa Oldireva Gustafsson's study of "The passive in nineteenth-century scientific writing" (pp. 110–135) which reveals that the shift to a de-personalized style in scientific writing characterized by a high frequency of passive constructions must have taken place much earlier than in the nineteenth century, despite previous claims to the contrary. While the individual authors sampled in the science genres of CONCE display considerable variation with regard to the use of passive constructions, the overall frequency and preferred verbs remain stable and very similar to PDE usage.

Christine Johansson examines "Relativizers in nineteenth-century English" (pp. 136–182) based on the categories science, trials and letters from the first and the third time period represented in CONCE, with regard to the distribution of *wh*-forms and *that* according to text type, restrictive vs. non-restrictive relative clauses or antecedent type and form, among other things. Surprisingly, the *wh*-forms predominate in all text types and environments, but especially in the scientific texts, with a frequency of roughly 90 per cent. The PDE development towards a predominant use of *that* had not yet set in, as nineteenth-century writers appear to greatly prefer the more explicit *wh*-forms.

The third paper with a focus on stability is Mark Kaunisto's study of "Anaphoric reference in the nineteenth century: *that/those + of* constructions" (pp. 183–193), e.g. *Lawrence's situation was almost as difficult as that of his brother.* Such constructions are today felt to be very formal. Kaunisto found the construction to be well-established in the CONCE data, but considerably more frequent in the texts from the science and debate categories. He also suggests a correlation between the use of anaphoric *that/those + of* and the number of words intervening between the pronoun and its referent. The construction appears more often in texts with a higher degree of anaphoric distance, which is also linked to formal writing.

The two remaining contributions cannot be classified according to the categories 'stability' and 'change'. Erik Smitterberg's in-depth analysis of partitives (partitive noun + *of* + prepositional complement) (pp. 242–273) shows that there is little diachronic change with regard to this construction, but instead a high degree of variation according to text type, with the highest frequencies in informative writing. Verb concord with partitive constructions also varies considerably, in terms of various semantic and syntactic factors.

Finally, Tony Fairman's contribution on "Words in English Record Office documents of the early 1800s" (pp. 56–88) stands out as it is not concerned with the morphosyntax of nineteenth century English, but rather with literacy, spelling and the teaching of spelling, especially with regard to the so-called 'lower

orders', i.e. artisans, shopkeepers and household servants, among others. Fairman analyses a corpus of Record Office documents, such as bills, pauper letters, baptism and marriage certificates, indicating differences in writing strategies due to different degrees of schooling. In comparing letters by less educated writers with those of well-educated writers, Fairman points out that only the latter group uses words of Latinate origin regularly and accurately. The analysis suggests that a great deal of research is still needed with regard to regional and social variation in nineteenth-century English. However, unlike the other contributors to this volume, Fairman did not rely on a representative corpus, but rather on a set of data compiled with the aim of collecting unusual or non-standard writing – which means a greater degree of bias in his data.

While some of the constructions under analysis are rather marginal, the contributions of this volume nevertheless offer tantalizing glimpses of nineteenth-century English. It is to be hoped that they inspire further corpus-based research on this period, perhaps even the compilation of additional corpora. It should have become clear from the studies in the present collection that the nineteenth century merits scholarly attention as it often provides the link between EME and PDE. Desiderata for further research include – as the editors of the present volume already point out in their introduction – regional and social variation in nineteenth-century English, as well as the development of the budding overseas varieties in New Zealand or Australia.

Note

1. A full list of texts for CONCE is given in the Appendix, pp. 272–277.

References

Bailey, Richard W. 1996. *Nineteenth-century English.* Ann Arbor: University of Michigan Press.

Bauer, Laurie. 1994. *Watching English change. An introduction to the study of linguistic change in Standard English in the twentieth century.* Edinburgh: University of Edinburgh Press.

Beal, Joan C. 2004. *English in modern times*: *1700–1945.* London: Arnold.

Görlach, Manfred. 1999. *English in nineteenth-century England: An introduction.* Cambridge: Cambridge University Press.

Mair, Christian. 2006. *Twentieth-century English.* Cambridge: Cambridge University Press.

Smitterberg, Erik. 2005. *The progressive in 19th-century English: A process of integration.* Amsterdam and New York: Rodopi.

Tony McEnery, Richard Xiao and **Yukio Tono**. *Corpus-based language studies: An advanced resource book.* London and New York: Routledge, 2006. 408 pp. ISBN: 978-0-415-28623-7. Reviewed by **Bernard De Clerck**, University of Ghent.

Provided it works and provided you are not an Eskimo, a refrigerator is a great invention. The logic behind this simple (or simplistic) observation might also be applied when reviewing a book: first of all does it prove to be useful for the target audience and secondly, does it 'work'? This review will revolve around answering these basic questions.

The book itself is part of the Routledge Applied Linguistics Series, whose target audience the series editors identify as "upper undergraduates and postgraduates on language, applied linguistics and communication studies programmes as well as teachers and researchers in professional development and distance-learning programmes" (p. xvi). The actual aim of the book is "to bring readers up to date with the latest developments in corpus-based language studies" by addressing both "how to" and "why" questions. The template that is used to realise this purpose is one that recurs throughout the series as a whole: an introductory part which explains key terms and concepts, an extension part which digs deeper by assessing and commenting on excerpts from selected key articles, and an exploratory section which puts theory into practice in student-oriented case studies and suggestions for further research. In the following paragraphs, I will first of all provide a concise summary of the material that is covered in each of the three parts. Secondly, and perhaps more importantly, I will provide personal comments on the content itself, the way it is conceptualised and its effectiveness in terms of the goals it wants to achieve.

In the first chapter of the introductory section the bare basics of corpus linguistics are covered by answering essential questions such as "what is a corpus?", "why use a corpus to study language?" and whether corpus linguistics is actually a theory or a methodology. The answers to these questions are both concise and insightful. They also nicely sketch a range of debates that has taken place against the background of these central issues. The authors diplomatically take a stand as well (by favouring corpus-based approaches and treating corpus linguistics as a methodology) though not without pointing out overlap between and justification for both approaches. Next, a number of important key concepts are introduced and discussed in a pedagogically justified order which is very similar to the stages one goes through when building and/or using a corpus and the questions and issues that are raised during the process. First of all, the impor-

tance of crucial concepts such as representativeness, sampling and balance are given centre stage and practical instructions are given on how these can be achieved (as far as a corpus can of course be truly representative). Unit A3 provides an overview of kinds of information that can be added to the raw text material, such as mark-up, POS-tagging, pragmatic and stylistic annotation, actual parsing and alignment in the case of multilingual corpora. Students and teachers will definitely welcome the distinctions that are made between these different kinds of annotation and their relative importance in terms of the research questions one is asking. It will help them (and researchers in general) to make the right choices in selecting existing corpora or accurately tagging one's own collected text material. Furthermore, attention is paid to the importance of statistics in corpus linguistics and to the different kinds of possible corpora that can be used. The recurrent pedagogical concern about terminological confusion is also very much reflected in the book's active concern (one that is much appreciated) with defining and differentiating the different kinds of labels, terms and kinds of corpora from one another (e.g. the distinction that is made between parallel corpora, comparable and comparative corpora, development corpora and learner corpora, etc.). Unit A7 provides an overview of some of the major publicly available "off the peg" – mostly English – corpora. Reference is not only made to widely known available corpora such as the BNC and the diachronic Helsinki corpus, but also to little gems such as the SED (Survey of English dialects). While of course not all corpora could be covered, reference is made (on a number of occasions) to the authors' companion website for a more comprehensive survey of well-known and influential corpora for English and other languages. Units A8 and A9 are particularly interesting for people who want to build their own corpus. Advice is given on how to extract usable data from the Web with the right corpus-processing tools (e.g. Grab-a-site, HTTrack, WebGetter, MLCT) together with warnings about copyright issues and how to clear them. In Unit 10, the concluding section to the introductory part and in my view one of the most stimulating chapters of the book, we are presented with an overview of corpus linguistics being used – more or less convincingly – in a number of areas of linguistics, including obvious domains such as lexicographic and lexical studies (with the invaluable import of corpus data in the study of collocations, semantic prosody and preference), grammatical studies, studies on register variation and genre analysis, contrastive, translation and diachronic studies, and studies on language learning and teaching. In addition, reference is made to work being done in the field of semantics, pragmatics, sociolinguistics, discourse analysis and forensic linguistics (with the intriguing case of Derek Bentley found innocent on the basis of linguistic evidence after being wrongfully

hanged in 1953). It is also worth mentioning that at the end of this chapter the authors do not shy away from pointing out the limitations in the use of corpora as well.

Section B, as noted above, is basically composed of excerpts from published material, which elaborate on and provide further background to the key concepts provided in Section A and related points of debate. Part 1 "Important and controversial issues" gives further support to the claim earlier made that external (or situational, social or extra-linguistic) criteria rather than internal (or linguistic) criteria should be used in initial corpus design by drawing upon two highly relevant works, namely Biber's (1993) "Representativeness in corpus design" and Atkins *et al*. (1992) "Corpus design criteria". These articles also foreground the related importance of stratified sampling both in terms of language production and perception. In addition, the reader can enjoy part of a very lively debate on the controversial issue regarding the role of corpora in linguistic analysis, language teaching and learning in excerpts taken from Henry Widdowson, Michael Stubbs and John Sinclair. As the excerpts point out, their viewpoints were or are in fact not that diametrically opposed as one (especially the authors themselves) expected or suspected them to be.

Units B3 to B6 present and illustrate some of the studies in the different fields of linguistics that have been introduced and illustrated in A10. More specifically, the use of corpora and corpus analysis is illustrated in lexical studies on the basis of excerpts taken from Krishnamurthy and Partington on collocation and semantic prosody respectively, which provide background knowledge for Case Study 1 in Section C. Grammatical studies such as Carter and McCarthy's account of the English *get*-passives in spoken discourse and Kreyer's study of genitive and *of*-construction in written English pave the way for Case Study 2 on the syntactic conditions which influence the choice between a *to*-infinitive and a bare infinitive following *help*. On the topic of language variation, studies are presented by Hyland and Kachru, who focus on metadiscourse in different scientific disciplines and definite reference in world Englishes respectively, and by Lehmann, who presents an analysis of subject relatives with a zero relativiser in American and British English. A more challenging and fairly complex study on register and genre variation is presented in Biber's multifeature/multidimensional (MF/MD) analysis, which is taken up again in the Exploration section as one of the most labour-intensive corpus-based studies. Contrastive and diachronic studies are represented by McEnery, Xiao and Mo's cross-linguistic study of aspect markers and by Kilpiö who traces the developments in the functions of the verb *be* from Old English to Early Modern English. Mair, Hundt, Leech and Smith in their turn report on shifts in part-of-speech based on the frequencies in

the matching LOB and FLOB-reference corpora. Contributions of corpus-based language studies to the field of language learning and teaching are presented in extracts from Gavioli & Aston,Thurstun & Candlin, and Conrad. These studies show the possibilities and limitations of real language data for language learning purposes and make clear that while corpora do not automatically guide us in deciding what should be taught, they can help us to make better-informed decisions and oblige us to motivate those decisions more carefully.

In the last section of the book, Section C "Exploration", McEnery *et al*. offer the reader the chance to carry out corpus-based analysis in case studies which are thematically linked to the A and B sections of the book. Not only do the authors present a step-by-step manual on how to carry out the searches themselves in view of the particular research questions, they also nicely foreground possible pitfalls in analysing results and doing statistics. In this way, the reader is taught the basic steps in operating the Concord and Keyword functions of the corpus-processing tool WordSmith, practical uses of the BNCWeb, as well as MonoConc Pro and ParaConc and the commonly used statistics package SPSS. At the end of each case study, readers are given further tasks to gain first-hand experience in using the tools and techniques just learned to solve language problems.

I greatly appreciated the book's fusion of theory, practice, technical knowledge and background reading. These, to me, are the most important ingredients for stimulating corpus linguistics research and having it carried out in a correct way by the target audience. Even if some of the key issues covered in the book may be common knowledge to the die-hard corpus linguist (who may perhaps be regarded as the Eskimo assessing the qualities of a fridge), they are nevertheless brought to the actual intended audience of the book, in a very "refreshing" manner, introducing them to or reminding them of lively debates which are stimulating both for laymen and experts. In addition, I particularly welcomed the many references for further reading which, at the time of publication, covered many of the most recently available studies and developments in tagging and data gathering.

In this way, this book not only puts corpus linguistics in the limelight as a very interesting way of carrying out linguistically relevant research, it also foregrounds the various disciplines in which it is used and stimulates the reader to think about related issues and to formulate other interesting research questions in the field of lexical studies, grammar, sociolinguistics etc.

On a more general level, the introduction-extension-exploration template is obviously a very practical and fruitful way of introducing and teaching corpus linguistics in the classroom. The introduction can pave the way to the students'

own reading and critical evaluation of the – preferably entire – articles in the extension section, whereas the exploratory section allows practical application and provides a stimulus for further experimentation and practice.

I will now address some minor points of criticism. First of all, while the authors stress the importance of representativeness, balance and sampling, it is only at a later stage that they acknowledge that attaining representativeness is not always feasible in practice. Not only do issues of copyright – which are in fact very briefly discussed – limit the possibilities or goals one has in mind, the very nature of the data itself seriously affects the size and diversity of the data one can process. One only needs to imagine the vast amount of spoken data that is produced at this very instance by native and non-native speakers of English to realise how underrepresented spoken data is in actual corpora. While the BNC is presented as a balanced corpus in Unit A2 (p. 17), the authors do not, at that point, address the imbalance between spoken (10%) and written data (90%). It is not until the section on DIY that the authors acknowledge that "[i]t is also important to note that the lower proportion of spoken data in corpora such as the BNC does not mean that spoken language is less important or less widespread than written language. This is simply so because spoken data are more difficult and expensive to capture than written data. Corpus building is of necessity a marriage of perfection and pragmatism" (p. 73). To be honest, pragmatism often gets the upper hand out of sheer necessity, a point which the authors could have made earlier.

Secondly, although of course not all recent developments or recently built corpora can be mentioned – as the authors themselves are the first to admit – I miss references to important projects which are aimed at taming the Web (GlossaNet and WebCorp, for instance) alongside the tools that are mentioned to retrieve web data in the overview section in Unit A7. In addition, while tools are presented to retrieve Web-based data, the authors themselves do not stress the inherent danger in using web data for linguistic purposes, which in view of the target audience might have been a useful reminder. Apart from obvious advantages of web data (its being freely available and constantly updated and fed with new material – not subject to the same delays in the creation of designed corpora), there are obvious disadvantages as well, such as the abundance of errors, made by both native and non-native speakers and the fact that the source of the data cannot always be traced. Additionally, using frequency data from a search engine is much more problematic than corpus-based frequencies, which seriously affects the validity of quantitative statements, the application of statistics and reliability in terms of representativeness and balance. See for example

Brekke (2000), Lawrence and Giles (1998), Meyer *et al.* (2003) and Renouf (2003) for more pros and cons of internet data.

Finally, a brief comment with respect to the case study on swearwords. The aim of this case study is to demonstrate the use of corpora in sociolinguistic studies and language variation by exploring differences in spoken and written registers based on sociolinguistic variables such as gender, age and social class. While the study itself shows a statistically significant difference in the use of swearwords (i.e. their frequency) for many of these parameters, it runs the risk of oversimplification. First of all, the output of an informant/informants is clearly not determined by one sociolinguistic variable at the time, but by the combination of these variables: they are of a certain age, belong to a certain social class, have followed a particular kind of education and are either male or female. In my view therefore, observations about language with respect to one variable can only be made if the others are kept constant. Now, even though the authors do combine some of the parameters, the data is not extensive enough to combine all and achieve statistical significance at the same time. Secondly, one's linguistic output is not only determined by one's own specific sociolinguistic parameters, but it is also influenced by those of the interlocutors. In fact, the analysis of the parameter 'intended audience' in written language showed significant quantificational differences between the use of swearwords for an all male intended audience and the use of swearwords for an all female intended audience (p. 282). The authors, however, do not transpose this finding to the results of the spoken data in which such a parameter is clearly operative as well. Whom one is talking to – male, female, young, old, education level and social class and the presence or absence of social distance – is at least as important as one's own sociolinguistic features, especially when it comes to using swearwords. This is one area where the results gained by corpus-based analysis should be positioned, interpreted and put into the perspective of a wider sociological context if one does not want to underemphasize the importance and complexity of the social dimension.

None of these minor flaws, however, diminishes the intrinsic value of this book in any serious way. It is a very fruitful marriage of theory, practice and up-to-date technical knowledge and a very useful course book which I would definitely consider using in teaching corpus linguistics. While the material covered may not shake the world of experienced corpus linguists (for whom it is not primarily intended in any case), this book is indeed a working refrigerator for anyone who wants to start teaching or doing corpus linguistics.

References

Atkins, Sue, Jeremy Clear and Nicholas Ostler. 1992. Corpus design criteria. *Literary and Linguistic Computing* 7: 1–16.

Biber, Douglas. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8: 243–257.

Brekke, Magnar. 2000. From the BNC toward the Cybercorpus: A quantum leap into chaos? In J.M. Kirk (ed.). *Corpora Galore: Analyses and techniques in describing English. Papers from the Nineteenth International Conference on English Language Research on Computerised Corpora* (Language and Computers 30), 227–247. Amsterdam and Atlanta: Rodopi.

Lawrence, Steve and C. Lee Giles. 1998. Searching the World Wide Web. *Science* 280: 98–100.

Meyer, Charles, Roger Grabowski, Hung-Yul Han, Konstantin Mantzouranis and Stephanie Moses. 2003. The World Wide Web as linguistic corpus. In P. Leistyna and C.F. Meyer (eds.). *Corpus analysis. Language structure and language use* (Language and Computers 46), 241–254. Amsterdam and New York: Rodopi.

Renouf, Antoinette. 2003. WebCorp: Providing a renewable data source for corpus linguists. In S. Granger and S. Petch-Tyson (eds.). *Extending the scope of corpus-based research. New applications, new challenges* (Language and Computers 48), 39–58. Amsterdam and New York: Rodopi.

Wolfgang Teubert (ed.). *Text corpora and multilingual lexicography* (Benjamins Current Topics 8). Amsterdam and Philadelphia: Benjamins, 2007. ix + 159 pp. ISBN 978-90-272-2238-1. Reviewed by **Christer Geisler**, Uppsala University.

Parallel corpora such as the English-Norwegian parallel corpus are by now well-established, but surprisingly few such corpora have actually been exploited as a source of information in the compilation of bilingual dictionaries. The present book brings together research from one much needed area: the use of corpus linguistic methods in bilingual and multilingual lexicography. It comprises a short preface, twelve articles, and an index. All contributors to the volume partici-

pated in the EU-funded TELRI project (*Trans-European Language Resources Infrastructure*). One outcome of the project was a multilingual parallel corpus with some dozen translations of Plato's *Republic*, many of which were aligned at the sentence level. The volume is a re-publication of a special issue of the *International Journal of Corpus Linguistics 6* (2001).

As is customary in edited volumes, the articles appear in alphabetical order according to the last names of the author. However, the work would have benefited greatly from a slightly different arrangement of the contributions. Wolfgang Teubert's article is the natural introduction to the field, and should have been the first article in the book, followed by John M. Sinclair's article on the COBUILD series of dictionaries and the concept of 'bridge dictionaries' (see below). Rūta Marcinkevičiene[1] discusses a number of crucial concepts, and this article should have been the third in the book. With these three articles the reader would have had a much better grasp of the field. As the use of corpus data in the field of bilingual and multilingual lexicography is still in its infancy, it would have been important to provide readers with some sort of road map.

Most of the articles report on research on parallel corpora and their use in lexicography/lexicology. R. Rossini Favretti, F. Tamburini, and E. Martelli investigate English and Italian legal terminology in the Bononia Legal Corpus (BOLC). In one case study of theirs, the Italian word *contratto* is compared to the English word *contract*. The authors argue that the collocations associated with various legal terms play a central role in the definition of their meaning, mirroring one of Wolfgang Teubert's arguments that the context of a word determines its meaning. Martin Čmejrek and Jan Curín analyse the results of automatic paragraph and sentence alignment as well as automatic extraction of a translation lexicon in an English-Czech parallel corpus. They report that, in a sub-corpus of messages from a computer operating system, the success rate is high, as opposed to a sub-corpus of texts from *Reader's Digest*. Hana Skoumalová reports on so-called bridge dictionaries, in this case, a partial translation of the English COBUILD dictionaries into Czech and Lithuanian. Since the electronic versions of such bridge dictionaries could be searched in any one direction, very special challenges face lexicographers. The compilation of the Croatian-English parallel corpus is treated by Marko Tadić and the compilation of the Russian-Finnish parallel corpus is dealt with by Mihail Mihailov and Hannu Tommola. The Croatian-English parallel corpus is based on newspaper texts, and the Russian-Finnish parallel corpus uses Russian classical fiction texts as source texts. Both studies discuss the problem of representativeness in using only a narrow range of registers in a parallel corpus. This is also mentioned by Martin Čmejrek and Jan Curín. L2 error analysis of corpus data is described in

an interesting article by Rafal Uzar and Jacek Waliński: translations by Polish students into English were annotated and marked for various types of errors, and this corpus of student translations was subsequently reused in translation classes. Tamás Váradi and Gábor Kiss focus on the problem of equivalence/non-equivalence, namely the fact that a word in the source language (English) frequently corresponds to a multi-word unit in the target language (Hungarian). In addition, they analyze the Hungarian translation equivalents of the English word *head*, showing that the range of equivalents is much broader in Hungarian compared with the translation equivalents of Hungarian *fej* 'head' in English.

The book was apparently printed in its present form without any changes. One wonders whether the authors should not have been given a chance to update their articles to take into account recent developments in the field. Another consequence of the articles having been reprinted in their original form is that, unfortunately, some editorial errors remain. For example, one article discusses the tax systems of "England" (p. 23), rather than "Britain". The same article refers to the concept of "internationalition" (p. 17) where the more felicitous word "internationalization" was probably intended. The word *occurrence(s)* is misspelled as "ocurrences" (p. 56) and "occurence" (p. 65). The very acronym ICAME is misspelled as "IACME" (p. 110). Several articles contain detailed descriptions of computer hardware specifics (such as type of network) and most of these descriptions are now either outdated or seem irrelevant. One article contains a list of references with fifteen works never cited in the text. Two articles have been published twice before.

Nevertheless, overall the studies in the volume are of value as they stress the benefits of including data from parallel corpora in future bilingual/multilingual dictionaries. In this context, as the authors emphasize, electronic versions of bilingual/multilingual dictionaries would have an advantage, as it would be impossible for printed dictionaries to contain large numbers of authentic examples in more than one language. Furthermore, the volume serves as a good survey of what uses can be made of parallel corpora in general.

Note

1. The typesetting software used did not reproduce all special characters in the authors' names as they appear in the book.