Hans Lindquist. **Corpus Linguistics and the Description of English**. Edinburgh: Edinburgh University Press, 2009. 219 pp. ISBN 978 07486 2614 4 (hardback); ISBN 978 0 7486 26151 (paperback). Price: (hardback) £ 60, (paperback) £ 18.99.

As the author points out in the opening paragraph of the first chapter of *Corpus Linguistics and the Description of English*, corpus linguistics is different from other "hyphenated branches" of linguistics, like sociolinguistics and neuro-linguistics, in that the former refers to a methodology for studying language, unlike the latter, where the names indicate what aspect of language is studied (language in society and language and the brain respectively). The title of the book indicates that in addition to dealing with corpus linguistics, the approach taken by the author is descriptive, i.e. an important underlying assumption is that by collecting and analyzing instances of authentic language use (in this case for English), important insights can be gained into different aspects of language (see Chapman 2006 for a discussion of the theoretical underpinnings of descriptive linguistics). The book is one in the series *Edinburgh Textbooks on the English Language – Advanced* and the intended readership is university students at the intermediate and advanced level, "who have a certain background in grammar and linguistics, but who have not had the opportunity to use computer corpora to any great extent in their studies" (p. xvi). The book thus aims to give the novice corpus linguist an introduction to the many aspects of language that can be studied by means of electronic corpora. Lindquist says on p. 1f. that "the ultimate aim is to learn more about the language and understand better how it works". There are study questions as well as a few suggestions for further reading at the end of each chapter, and in addition there is an accompanying web page with links to online corpora and hands-on exercises for each of the ten chapters. The author points out in the preface that the book is about *using* corpora – not about creating one's own (although in the final chapter there are some hints as to how one can go about compiling corpora from the web).

The presentation is typical of a certain kind of textbook aimed at students, with the reader frequently being addressed as *you*. The first person is used extensively (more often inclusive *we* than *I*), and imperatives (*note that…*) are occasionally used when step-by-step concrete examples are given. For someone who is *not* a student (like me), this interpersonal style may come across as a little condescending at times (the teacher is leading the student by the hand, so to speak), but overall, the book has so many good qualities that it is easy to disregard this aspect of it and instead be drawn into the many interesting applications of corpus linguistics that the author presents in the volume.

The book contains ten chapters. The first two give a general background of corpus linguistics, and the following eight chapters, each roughly 20 pages in length, deal with specific areas of English, such as lexis, grammar, and gender in language. There is some overlapping information in the last eight chapters; however, this is an advantage rather than a drawback, as students may begin by reading the first two chapters and then select the areas they are interested in and

focus directly on these, skipping over others that are of less interest. With a few exceptions, the book is written in a very accessible way and really does not assume prior knowledge of corpus linguistics, which makes it an excellent starting point for anyone wanting to find out more about – and test – the many applications of corpus linguistics in language research.

Chapter 1 gives a historical survey of the use of corpora in linguistics and explains some important concepts, like concordance and frequency. Arguments for and against the use of corpora in language studies are also presented, and the chapter ends with a comprehensive survey of corpora, many of which can be accessed via the web. The author does not include any web addresses for the corpora, which is probably a wise thing to do, as such addresses tend to become obsolete over time. Instead, the accompanying web page has links to a number of on-line corpora. The advantage of doing it this way is of course that the web page can be updated whenever an address is changed.

The second chapter is on "counting, calculating and annotating". Here, important basic concepts like POS (part of speech) tags, lemmas, and type vs. token are explained. The reader also finds out about simple statistics, like how to use percentages and how to normalise frequencies when comparing corpora of different sizes ('normalising' means to give the number of occurrences not as absolute frequencies but per a predetermined number of words, e.g. 'per million words'). The many tables give the reader a good idea of the various ways of presenting information on frequency. The section on part-of-speech tagging is very informative and contains a list of the most commonly used tags. This part also raises the problem of incorrect tagging (automatic tagging will inevitably result in a number of incorrect tags being assigned), and the examples of such incorrect tagging, presented in Figures 2.1 and 2.2, illustrate the problems well. There is also a section on how to use chi-square to test the statistical significance of one's results.

The following eight chapters focus on specific areas of study and as already mentioned, there is a certain amount of overlap between them. For instance, there is a separate chapter on language change (chapter 9), despite the fact that many of the illustrative examples given in previous chapters *are* of language change. Chapters 3–5 all present aspects of how words work in English but from slightly different perspectives (lexis in chapter 3, collocations in chapter 4 and phraseology in chapter 5). Chapter 6 is on metaphor and metonymy. However, already in chapter 3, which deals with words and their meanings, it is shown how metaphorical meanings are often more common than the literal meaning for certain words in English (p. 54). Chapter 7, on grammar, overlaps with the section on colligations in chapter 4 and the illustration of POS-grams in chapter 5. Chapter 8, which has a sociolinguistic slant, presents ways of studying gender and language using corpora. Finally, chapter 10 suggests various ways of using the web as corpus. It is not obvious to me what principles were used in the choice of topics or in deciding on the order of presentation. Nevertheless, I believe these

chapters function both to give an overview of various aspects of language that can be studied using corpora and to suggest specific topics for student papers.

Chapter 3, on lexis, begins with a presentation of lexicography before and after the introduction of electronic corpora. Lindquist demonstrates how corpora can be used to find out more about words and their meaning than what can be had from dictionaries. He goes on to present some important concepts in the study of lexis, namely collocation (words that frequently co-occur), colligation (the grammatical categories that a word co-occurs with) and semantic prosody (the way meanings spread over several words, often conveying negative, less often positive evaluation). Changes in lexis over time are demonstrated by showing how *greenhouse effect* has given way to *global warming* and how *maybe* is gaining in frequency at the expense of *perhaps*. There are also illustrative examples of how words spread between varieties of English and on the language of literature (literary stylistics). The illustrations of the latter – from Michael Stubbs' study of Joseph Conrad's *The Heart of Darkness* and Jonathan Culpeper's investigation of the language of the six main characters in Shakespeare's *Romeo and Juliet* – show that corpus studies may have a lot to offer not only to linguists but also to literature students.

Chapter 4 is called 'Checking collocations and colligations'; however, out of the 20 pages, only a page and a half at the very end describes work on colligations, while the rest of the chapter is devoted to collocations. The first part of the chapter contains information on two statistical measures for identifying words which frequently occur in the vicinity of another word (usually within a span of four or five words on either side of the word in question), namely the mutual information (MI) score and the z-score. The second part of the chapter is on 'adjacent collocations', i.e., words which occur immediately before or after some other word(s). This is illustrated in tables which give the most common right collocates of *go to* (the most frequent ones being *bed* and *sleep*), three-word combinations ending in *bed* (with *on the bed* topping the list and *go to bed* only in third place), and left collocates of the noun *hand* (where *other hand* by far outnumbers other combinations) in British English, taken from the freely available *Phrases in English* database (see below). This second part ends with a fairly comprehensive account of Lindquist's own investigation of the syntax and semantics of *at/in/on/to hand*. As already mentioned, colligations are treated only summarily in a short section at the end of the chapter.

Chapter 5 is on the study of phrases, i.e. "more or less fixed strings which are used over and over again" (p. 91). Lindquist opts to use the term *phrases* rather than some of the other terms that proliferate: 'idioms', 'fixed phrases', 'recurring strings', 'formulaic sequence' etc. The problem of differentiating between such phrases and the adjacent collocations dealt with in the previous chapter is clear from Lindquist's definition of adjacent collocations as "a more or less fixed string, sometimes with an open slot where there is some variation" (p. 71), something which the author also acknowledges on p. 91, where he says that "phrases are *close to* the second type of collocation …, namely adjacent

collocations" (my italics). The first part of the chapter, on idioms, contains a very interesting example of how the idiom *storm in a teacup* can be exploited for creative purposes (like *storm in a whiskey glass*) in different varieties of English. Here, the author has used the Google advanced search mode, and set the domain to e.g. uk (for British web pages) or nz (for New Zealand). The second part of the chapter describes how n-grams can be studied in English. An n-gram is a recurring string of identical word sequences, where n stands for any number; researchers usually study n-grams of between two and eight words which occur a minimum number of times in a corpus. Lindquist makes use of William Fletcher's database *Phrases in English* (PIE), based on the one hundred million-word British National Corpus (BNC), which lists all n-grams occurring three times or more in the corpus. As a start, the most frequent 5-grams in the BNC are presented, followed by examples of how one can find out more about when and how these 5-grams are used (examples include concordances of *at the end of the* and *I don't know what*). There is also an illustration of how POS-grams can be studied (part of which is a repetition of the information on colligations at the end of chapter 4). The final section shows how n-grams can profitably be applied in literary analysis; here, Mahlberg's 2007 study of recurrent phrases in Dickens is used as an example.

Chapter 6, on metaphor and metonymy, begins by giving a general introduction to the area of study, and has a very informative section on the difference between creative metaphors, conventional metaphors and dead metaphors, followed by a short section on similes. There is also a very short introduction to metonymy (p. 118f.), including a useful table illustrating three types of metonymy. Unfortunately no suggestions as to how metonymy can be studied are given. Lindquist then goes on to discuss conceptual metaphor as set out in Lakoff and Johnson (1980). The main problem of studying conceptual metaphors in a corpus has to do with the fact that they are schematic and underlie linguistic metaphor. For instance, an expression like *this relationship is a dead end street* shows that we conceptualize love as journey. The investigator therefore needs to identify such linguistic metaphors *before* a conceptual metaphor can be studied in a corpus, which is not a straightforward matter. Lindquist takes as an example the conceptual metaphor THE MIND IS A MACHINE and shows how it can be studied by means of checking how often phrases such as e.g. *a bit rusty* and *well-oiled* refer to thinking in a corpus. Abortive searches are accounted for, which I find to be one of the attractions of the book, as it underscores the fact that using corpora is not always simple, and that even experts (like Lindquist) will often have to test various ways of finding material of interest for a particular study. Finally, there is a section on how manual analysis might be a good starting-point for corpus studies of metaphor.

Chapter 7, on grammar, begins with three diachronic studies of language change in English, namely the frequency of *who* vs. *whom*, the increase of *get*-passives in 20th century English and finally how the infinitive has given way to the *–ing* form after *accustomed to* in the past two hundred years. A detailed

presentation of a synchronic study of the frequency and use of *difficulty (in) V-ing* (here V stands for any verb) is given at the end of the chapter. It is often difficult to get at various grammatical forms in untagged corpora, and the detailed presentation of what search strings can be used is, therefore, welcome. If the search is too general, too many irrelevant cases may show up; if it is too specific, the number of relevant hits may be too low. Lindquist also shows very convincingly how earlier studies can be used as the starting point for extended searches, which in their turn can lead to discoveries which expand our knowledge of how language works.

Chapter 8, called "Male and female", looks at ways of studying gender and language by means of corpora. The first section deals with the changing words for the professional roles of men and women (from *firemen* to *firefighters*, for instance). The second section suggests how one can find out about how men and women are portrayed in texts by looking at what adjectives are used to modify *man* and *woman* in a corpus and what verbs are used when *she* is the subject and *him* the object, as opposed to when *he* is the subject and *her* the object. Finally, in corpora which contain information on the gender of the speaker (like the BNC and MICASE, the Michigan Corpus of Academic Spoken English), it is possible to study how women and men *use* language. The final section of the chapter is, accordingly, on men's and women's use of hedging, i.e. phrases such as *sort of* and *kind of*.

In the opening paragraph of chapter 9, on language change, Lindquist refers back to the many illustrations of language change that have already been presented in the book, and points out that this is "one of the most rewarding things you can study in corpora" (p. 167). In this chapter, the author presents two studies, one of changes in apparent time (a comparison of the use of *likely* as an adverb in British and American English) and one of changes in real time (the development of an abstract meaning of *beside(s)*) over the centuries. The former can be studied using corpora of present-day English; for the latter, historical corpora are needed. The chapter ends with a section on how the Oxford English Dictionary can be used for long-term diachronic studies. Table 9.6 (based on Mair 2004) is very useful in this respect, in that it contains information on the number of quotations and estimated number of words for each century from the year 1000 to the present day.

The final chapter is on "corpus linguistics in cyberspace" and shows how the web can be used *as* a corpus in cases where regular corpora are too small or dated, or no corpora containing the varieties one is interested in has been compiled, as well as how it can be used *for* a corpus, i.e. to compile one's own corpora. Lindquist does not fail to point out the instability of web pages, which makes it difficult to replicate studies, but nevertheless, the examples given in the areas of phraseology, regional variation and dialect/non-standard English serve as good models for small investigations, especially when one wants to find out more about low-frequency vocabulary and structures, where even large corpora might not contain enough material.

The accompanying web page has one page of exercises for each of the ten chapters (two or three exercises per chapter), in addition to direct links to the corpora and other resources used in the exercises. I tested a number of them, and found that with a few exceptions, they work quite well. The instructions are very clear and even novice corpus users should have no problem following the steps. Most of the exercises also yield interesting (and even somewhat surprising) results, at the same time as they are not too time-consuming to do. The one exercise that really does not work well is 6.1, on conceptual metaphor (*life/love is a ...*). One problem is that it takes a long time to go through the suggested 100 examples for each of the two searches; another is that there are many irrelevant hits, and the students will actually need to look at the context for quite a few of the concordance lines in order to see whether they really are metaphorical. Often, the following word is an adjective, which may or may not be modifying a generic noun. At other times, the following noun does not refer to an aspect of *life* or *love* but rather to a noun that comes earlier. For instance, the concordance line *love is a lie* turns out to be part of the sentence "the equalizing agent of this triumph of love is a lie". In any case, this is a minor shortcoming, and this exercise may still be useful, as it may lead to interesting discussions about metaphors. The exercises in general are varied and interesting and will give students a good insight into how to do corpus studies.

Although the book has many strong points, what I would have liked to see described in a slightly more accessible way are some of the more advanced statistical measures. I believe a student would have a hard time carrying out a chi-square test with the help of the information in the book – personally, I would refer them to Levon (2010: 77ff.) or Johannesson (1993: 93ff.) instead. For MI scores and z-scores, I would recommend students to look at chapter 4 in Hunston (2002). The treatment of n-grams is also a little confusing. The concept is explained in parenthesis on p. 81, and on p. 87, POS-grams are contrasted with n-grams, but it is only on p. 101 that n-grams are actually explained. The book is well edited on the whole, although there are a few typos, e.g. *lead* instead of *led* on the last line of p. 11, *try think* instead of *try to think* on p. 54, and curiously, in example (17) on p. 84 from the BNC, the ferret is said to be near the *whole* rather than the *hole* (this is correct in the corpus, however).

All in all, this is a book that is sorely needed, and one that I believe will work well both as a text book for a course on corpus linguistics and as a guide for students who are embarking on a corpus linguistic study for their degree paper in English. Indeed, the book is suitable for anyone wanting to learn how to carry out corpus investigations on their own. One thing that I like very much about the book is that many chapters begin with a brief survey of the historical development of the topic at hand. The reader is thus not led to believe that the use of corpora began with the invention of computers, even though admittedly, computers have made life a lot easier for corpus linguists. Lindquist is also careful to present critical voices and to weigh the pros and cons of various aspects of corpus linguistics. He presents many original examples of corpus studies as well as

research done by other linguists, often following them up with his own investigations of the same research questions, but using different corpora than the ones in the original study. Similarly, the studies presented in the book may serve as inspiration for students to embark on their own investigations. A problem that I have encountered many times is that it is hard for students to come up with original ideas of studies that can be done using corpora. Still, the many interesting step-by-step investigations presented in the chapters and suggested in the exercises might make this book just the resource that is needed to spread the knowledge of how to use corpora and make them tools that can be used not only for linguistic research, but also for finding information that cannot easily be gleaned from reference books.

Solveig Granath

References

Chapman, Siobhan (2006). *Thinking about Language: Theories of English*. Houndsmills, Basingstoke, Hampshire: Palgrave macmillan.

Hunston, Susan (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Johannesson, Nils-Lennart (1993). *English Language Essays: Investigation Method and Writing Strategies*. Fourth edition.

Lakoff, George & Mark Johnson (1980). *Metaphors we live by*. Chicago: Chicago University Press.

Levon, Erez (2010). Organizing and processing your data: The nuts and bolts of quantitative analyses. In L. Litosseliti (ed.), *Research Methods in Linguistics*. London: Continuum, 68–92.

Mair, Christian (2004). Corpus linguistics and grammaticalization theory: Statistics, frequencies and beyond. In H. Lindquist & C. Mair (eds.), *Corpus Approaches to Grammaticalization in English*. Amsterdam: Benjamins, 121–150.

Mahlberg, Michaela (2007). Corpus stylistics: Bridging the gap between linguistics and literary studies. In M. Hoey, M. Mahlberg, M. Stubbs & W. Teubert, *Text, Discourse and Corpora*. London: Continuum, 219–246.