Reviews

Gunter Kaltenböck. It-*extraposition and non-extraposition in English. A study of syntax in spoken and written texts* (Austrian Studies in English 90). Vienna: Wilhelm Braumüller. 2004. 324 pp. ISBN 3-7003-1461-2. Reviewed by **Magnus Levin**, Växjö University.

Extraposition, a term which seems to have been coined by Jespersen (1949 III: 25), has received a great deal of attention in transformational analyses but only a rather limited number of functional studies have been devoted to the subject. Gunter Kaltenböck's very thorough functional analysis is therefore particularly welcome. The discussion throughout the book is richly illustrated with authentic examples and covers a number of factors influencing the choice between *it*-extraposition and non-extraposition, such as information structure, the principle of end-weight and register. The study clearly illustrates the advantages of a corpus-based approach when investigating the variation between the two alternatives. The use of naturally occurring data makes it possible to take the information structure into account, and Kaltenböck presents a great deal of evidence suggesting that the two alternatives are not generally interchangeable in authentic texts.

Typical text-book examples of extraposition can be seen in (1) and examples of non-extraposition in (2) (p. 1). In examples such as these it seems possible to use the two constructions interchangeably.

(1a) It is surprising *that John went to Paris*.
(1b) It is obvious *what John will be doing in Paris*.
(2a) *That John went to Paris* is surprising.
(2b) *What John will be doing in Paris* is obvious.

Apart from *that*-clauses and *wh*-clauses as illustrated above, the study also includes gerunds, *to*-infinitives, *for*/*to*-clauses and less prototypical instances such as *if*-clauses and *when*-clauses.

Chapter 1 of the book introduces the study, and Chapter 2 provides a selective overview of some previous work in the area. Chapter 3 is devoted to the complex task of delimiting the area of study. Chapters 4 and 5 constitute the main body of the book, the former discussing the formal properties, and the latter the functional properties of the constructions. The book ends with a conspectus of the factors influencing the choice of non-extraposition in Chapter 6 and a summary and conclusion in Chapter 7.

The material comes from the British component of the International Corpus of English (ICE-GB), which consists of 600,000 words of spoken and 400,000 words of written British English from 1990 to 1994. Kaltenböck groups the 'written to be spoken' texts with the written texts and thus ends up with two subcorpora of equal size. The material used is considerably larger than in previous studies, and this is particularly relevant since non-extrapositions are very much rarer than extrapositions, and the author can therefore carry out a detailed analysis also of non-extrapositions. Nevertheless, the predominance of extrapositions is also reflected in the presentation of the material where considerably more space is provided for extrapositions than non-extrapositions.

The discussion of previous studies in Chapter 2 focuses mainly on recent corpus studies, such as Mair (1990), but some of the most influential generative studies are also presented. Kaltenböck devotes considerable attention in Chapter 3 to a thorough and lucid discussion of the delimitation of the area because (non) extraposition has previously been delimited in different ways by different authors, and because there are several constructions that are similar to extraposition, such as *it*-clefts and right-dislocations. It could be argued, however, that some of the constructions included, such as *it seems/appears*, which disallow non-extraposition, could have been excluded from the material.

Chapter 4 provides a comprehensive discussion of the formal properties of *it*-extraposition and non-extraposition. The study shows that *it*-extraposition is clearly preferred over non-extraposition in both speech and writing (90.2% extraposition in speech and 87.6% in writing). This observation and the fact that English strongly favours light subjects lead the author to the conclusion that non-extraposition is the (statistically) marked alternative (cf. Givón's (1995) criteria for markedness). This runs counter to the more commonly held perception that non-extraposition is unmarked because it is more 'basic' in structure and has canonical word order. It should nevertheless be noted that it might be mis-

leading to rely too heavily on frequency as a criterion for markedness (see e.g. Pinker & Prince 1994).

The study discusses both subject extraposition and object extraposition, but since object extraposition, as in (3), is fairly marginal in the material (accounting for only 5.9% of the total number of extrapositions) the analysis is concentrated on subject extrapositions.

(3) The Norwegians find it tough *going through the thick jungle of ice* (S2B-024-83)

One of the points that Kaltenböck's material illustrates repeatedly is, as indicated above, that it is often doubtful whether extraposition and non-extraposition are really substitutable in authentic texts. For example, non-extraposition in (3) (?The Norwegians find *going through the thick jungle of ice* tough) is prevented because it violates the principle of end-weight. This lack of options is also illustrated by the fact that the different complement clauses in subject extrapositions have very different preferences: *that*-clauses, *to*-infinitives and *for*/*to*-constructions favour extraposition, while *ing*-clauses mainly occur in non-extrapositions, and *wh*-clauses are evenly distributed between extraposition and non-extraposition.

Chapter 5, which deals with the functional properties of (non-)extraposition, provides the most interesting findings. The author begins by presenting a detailed account of the various degrees of discourse familiarity and their correlation with (non-)extraposition. New (i.e. irretrievable) information is sub-divided into brand-new or new-anchored information, while given (i.e. retrievable) information is divided into inferrable (directly retrievable) and textually and situationally evoked. The material shows that there are distinct differences between extraposed and non-extraposed clauses as regards their degree of givenness. Extraposed clauses predominantly contain new information (71.5%), while non-extraposed clauses contain given information in 80.2% of the cases. There is more frequently new information in extraposed complement clauses in writing (83.2%) than in speech (56.1%).

The communicative functions of extraposed clauses with given information differ greatly from extraposed clauses with new information. For instance, extraposed given clauses are often of the 'reaction mode' type where the matrix clause expresses a certain opinion towards something previously stated, as in (4). New complement clauses, on the other hand, are often attached to a matrix predicate that sets up the relevance for the complement, as in (5).

(4) I've got William at home on this gap And *it is very nice* having William home on this gap (…) (S1A-031-2)

(5) 5.2.2 The physical environment of coastal sand dunes <p> First it is necessary *to give a brief outline of how coastal sand dunes form and evolve*. (W2A-022-59)

Examples such as these are representative of the rich evidence the author draws upon in his exhaustive analysis of the communicative functions. Another major functional issue is the distribution of weight in the sentence. In the ICE-GB material extraposed complement clauses are roughly three times longer than their matrix clauses. With non-extraposition, however, subject clauses and matrix clauses are of equal length. Kaltenböck therefore concludes that appropriate weight distribution is only of limited importance in non-extraposition, since it can be overridden by other factors, such as the distribution of information.

The last part of Chapter 5, which concerns the functional aspects of non-extraposed clauses, is especially interesting since it deals with the marked instances. Marked instances are often particularly revealing since they are not only important in themselves but also shed light on the typical features of unmarked instances, and therefore this part of Chapter 5 could have been even more extensive.

Kaltenböck distinguishes three main communicative functions of non-extraposition: a commenting function, a cohesive function and 'presenting new information as given'. The commenting function refers to cases where the matrix clause contains a comment on the topic in the subject clause. The cohesive function means that speakers adhere to the given-before-new principle whereby the retrievable information at the beginning of one sentence is linked to the new information at the end of the preceding sentence. Finally, 'presenting new information as given' is a rhetorical device which in some cases involves referring to 'shared' or 'general knowledge'. In other cases it may be used in persuasive discourse, as in (6), where a speaker tries to persuade the listeners by presenting the idea that the court's reasoning is wrong as a given fact (from a lecture (p. 276)).

(6) Implicit in the court's reasoning is the assumption that ownership is absolute or it's not ownership *That this is wrong* hardly I think needs demonstration (S2B-046-87)

In conclusion, this is an exhaustive investigation of an area that has received a great deal of attention in corpus studies in recent years. It is based on a carefully

sampled standard corpus, and it demonstrates in a convincing manner how formal and functional factors interact in the choice between the alternatives. That the understanding of the variation between extraposition and non-extraposition has increased greatly through Kaltenböck's work is unquestionable.

References

Givón, Talmy. 1995. *Functionalism and grammar.* Amsterdam & Philadelphia: Benjamins.

Jespersen, Otto. 1949. *A modern English grammar on historical principles* III. London: Allen & Unwin.

Mair, Christian. 1990. *Infinitival complement clauses. A study of syntax in discourse*. Cambridge: Cambridge University Press.

Pinker, Steven and Alan Prince. 1994. Regular and irregular morphology and the psychological status of rules of grammar. In Susan D. Lima, Roberta L. Corrigan and Gregory K. Iverson (eds.), *The reality of linguistic rules*, 321–351. Amsterdam: John Benjamins.

Hans Lindquist and Christian Mair (eds.). *Corpus approaches to grammaticalization in English.* Amsterdam and Philadelphia: John Benjamins, 2004**.** xiv + 264 pp**.** ISBN 90-272-2284-3, 1-58811-523-2. Reviewed by **Andrea Sand**, University of Hanover.

The volume *Corpus approaches to grammaticalization in English* is based on a selection of papers first presented at the international symposium on corpus research in grammaticalization in English held at Växjö University, Sweden, in 2001. The original contributions from the symposium have been complemented by further papers solicited from linguists with a strong research background in the area of corpus-based grammaticalization research. In their introduction, the editors, Hans Lindquist and Christian Mair, comment on the recent *rapprochement* in English linguistics between corpus linguists and grammaticalization theorists working on linguistic processes such as cliticization, semantic bleaching

or phonological reduction. They formulate the threefold aims of the volume, namely to make the point that the study of grammaticalization can benefit from the data and methodology developed within corpus linguistics, to suggest methodological refinements for the study of grammaticalization based on achievements in corpus linguistics and to present case studies of grammaticalization on the basis of rigorous data analysis, thus complementing more typologically oriented research dealing with less well-documented languages than English.

The nine papers can be broadly categorized into two groups, depending on their main focus. Although all contributions contain corpus-based analyses of grammaticalization processes, some have a stronger theoretical and methodological focus while others can be primarily regarded as case studies on the development of various aspects of English grammar, such as verb forms (e.g. the papers by Marianne Hundt or Laura Wright) or prepositions (e.g. the paper by Matti Rissanen).

In the first group, Terttu Nevalainen's paper also serves as a short introduction to the topic at large. She examines 'Three perspectives on grammaticalization: Lexico-grammar, corpora and historical sociolinguistics' (pp. 1–31), showing how each area contributes to our understanding of the processes at work. Nevalainen stresses the importance of distinguishing text-type specific variation from ongoing linguistic change and the insights gained by contextualizing linguistic data with the help of socio-historic information. With regard to the use of corpus linguistic methodology, she points out the problems of corpus annotation (especially with regard to word class tagging) and normalization (especially with regard to spelling variants) in the work with historical corpora. Her case study traces the development of the intensifiers *fair(ly)* and *prett(il)y* as examples of *-ly* adverbialization. Christian Mair argues forcefully in favour of an integration of 'Corpus linguistics and grammaticalisation theory: Statistics, frequencies, and beyond' (pp. 121–150). Using the *OED*'s quotation base as a tool for a long-term historical analysis, he proceeds to give a number of examples which illustrate two different grammaticalization patterns; one Mair labels 'dynamic type', which is characterized by a delayed-onset increase in frequency of the item in question (as in the case of the *going to*-future), and another that he calls 'static type', which occurs independently of statistically relevant frequency increases and can only be detected by means of a qualitative data analysis (as in the case of the complex conditional subordinator *supposing that*). Sebastian Hoffmann addresses the question 'Are low-frequency complex prepositions grammaticalized? On the limits of corpus data and the importance of intuition' (pp. 171–210), critically assessing the potential and limitations of corpus-based analysis with regard to low-frequency items, such as complex prepositions (e.g.

by dint of or *in proximity to*). Hoffmann uses a variety of sources, such as the *British National Corpus*, an electronic newspaper corpus, a database from Project Gutenberg and the quotation base of the *OED* on CD-ROM, to identify problems of corpus representativeness and statistical significance as the major obstacles in the study of low-frequency items. He proposes the concept of 'grammaticalization by analogy' (pp. 194f.) as a solution to this dilemma, arguing that low-frequency items may be treated according to the patterns used for high-frequency items in the mental lexicon, especially if a sequence has high saliency. Laurel Brinton discusses 'Subject clitics in English: A case of degrammaticalization?' (pp. 227–256), addressing the debate in grammaticalization theory on the unidirectionality of grammaticalization by looking at a supposed case of decliticization, namely the replacement of forms like *hastow* or *wiltow* by *hast thou* and *wilt thou*. Brinton's main argument is that these are not counterexamples to grammaticalization as such, as the full forms continued to exist and just replaced the cliticized forms again once the phonological rules governing their use had disappeared in Early Modern English.

The case studies presented in the second group of papers serve as illustrations of the synergetic combination of corpus linguistics and grammaticalization research. Sali Tagliamonte studies '*Have to, gotta, must*: Grammaticalisation, variation and specialization in English deontic modality' (pp. 33–55) in the York English Corpus. Using multivariate analysis, Tagliamonte describes the distribution of the deontic modals found in the York dialect and comes to the conclusion that the variety is rather conservative compared to other varieties of British or American English, as *got to* and *gotta* are still quite rare. Karin Aijmer analyses 'The semantic path from modality to aspect: *Be able to* in a cross-linguistic perspective' (pp. 57–78), comparing data from the English-Swedish Parallel Corpus with regard to English and Swedish expressions of ability. She comes to the conclusion that, while there are parallel tendencies in the use of success verbs (e.g. English *manage*) to express ability and the development of aspectualizers (e.g. Swedish *lyckas*), cross-linguistic generalizations are difficult, as the language-internal factors in the development play a primary role. Marianne Hundt investigates 'The passival and the progressive passive: A case study of layering in the English aspect and voice systems' (pp. 79–120), using the example of the progressive passive (e.g. *The house is being built.*) and its rival, the passival (e.g. *The house is building*.) as an illustration for the notion of stable layering as opposed to transitional layering, in which one element is gradually ousted by a new construction. In the case discussed by Hundt, the passival has become marked but has not completely disappeared, as her analysis of the ARCHER corpus shows. Matti Rissanen relies on a number of different corpora to study

'Grammaticalisation from side to side: On the development of *beside(s)*' (pp. 151–170), which goes back to Old English *be + sedan* and grammaticalized in the Middle English period. Rissanen's analysis proves that the development of *beside(s)* proceeded parallel to a number of other adverbial connectives, such as *before* or *until.* Finally, Laura Wright discusses 'Life after degrammaticalisation: Plural *be*' (pp. 211–226), challenging the notion that degrammaticalization automatically entails lexicalization by studying the uses of present plural *be* in her corpus of 16th and 17th century court proceedings. As Wright is able to show, plural *be* was not completely replaced by *are*, but rather survived by acquiring dialectal and ethnolectal properties (especially in the United States), resulting in layering in terms of sociolinguistic distribution and indexicality (cf. also the contribution by Marianne Hundt).

While the different contributors cannot seem to agree on whether their subject should be spelled 'grammaticalization' or 'grammaticalisation' (as the titles quoted in this review reveal), they all agree on the suitability of corpus linguistic methodology in grammaticalization research. This collection certainly makes a strong point on the advantages and insights gained from the synthesis of grammaticalization theory and corpus linguistic methodology, especially with regard to the adequate assessment of the role of frequency in grammaticalization and the need to combine quantitative and qualitative data analyses in order to gain further insights into grammaticalization and other diachronic processes. It is to be hoped that it will inspire new research along these lines.

Nadja Nesselhauf. *Collocations in a learner corpus.* Amsterdam and Philadelphia: John Benjamins, 2005**.** xii + 332 pp**.** ISBN 90-272-2285-1, 1-58811-524-0. Reviewed by **Göran Kjellmer**, University of Göteborg.

Collocations have become quite fashionable as a study object over the last decade or so. This is very largely a consequence of the creation of extensive electronic general language corpora, which have enabled us to study phenomena that could only be noticed and processed with the help of a computer. The advent of learner corpora has put yet a new tool in the hands of linguistic researchers, a tool that has been used for various purposes, not least the identification of overused and underused items in learner language. The author of the

book under review, Nadja Nesselhauf, makes use of a learner corpus in order to focus on collocations in a corpus of English produced by advanced German students and to study how the collocations are used, what infelicities are produced, what the reasons for the latter could be, and finally how deviant collocations could be foreseen and avoided.

Having presented her topic and briefly discussed previous research in the area, the author states her aims (p. 9), which are to identify typical difficulties of advanced learners in the production of collocations, to identify the factors that contribute to the difficulty of (certain) collocations, to find out what material and strategies learners use to create collocations, and to formulate suggestions for language teaching. The material on which the investigation is based comes from the German part of the ICLE corpus, with the restriction that only verb-noun combinations in argumentative essays are analysed, altogether over 2,000 instances.

Distinctions, definitions and classifications in the field of word sequences are notoriously troublesome, and the author puts a good deal of effort into establishing her own system with regard to verb + noun combinations. She distinguishes collocations (*shrug one's shoulders, make a decision*) from free combinations (*want a car, read the paper*) on the one hand and from idioms (*sweeten the pill, kick the bucket*) on the other. A verb-noun collocation is described thus:

> The noun can be used without arbitrary restriction in the sense in which it is being used, but the verb is, in the given sense, to some degree arbitrarily restricted to certain nouns. (p. 33)

To determine the acceptability of the sequences produced by the students a group of dictionaries were consulted, and in the cases where the sequences were unacceptable or possibly unacceptable they were referred to four judges, two British and two American, who pronounced judgement on them. The result of the subsequent calculations was that 69 per cent of the sequences produced were deemed clearly or largely acceptable, 1,432 out of 2,082 (p. 69). The major part of the rest of the discussion is then devoted to the 31 per cent that were questionable or unacceptable.

In Chapter 3, the longest chapter, a detailed analysis of deviating sequences leads to a fine-meshed classification of the material, where the number of occurrences in each category gives an indication of the degree of difficulty of the type in question. Deviations in the verb and the noun phrase are considered separately, as are more global deviations. Corrections and suggestions by the judges

form the basis of so-called "Collocations intended", which can then be compared with "Combinations produced". As is natural in a quantitative study like this, figures and tables abound.

Chapter 4 contains a discussion of the building material of deviant collocations, both that deriving from L2 (English) and that deriving from L1 (German). The identification and classification of such material is thought to have some prognostic value; the results will be made use of in the last chapter.

In Chapter 5, factors correlating with learners' difficulties with collocations are discussed, among them the extremely important factor of congruence or non-congruence between L1 and L2. Questions here are to what extent deviant collocations are due to similar or identical constructions in L1 (non-congruence between L1 and L2), and, in some measure, to what extent non-deviant collocations are due to congruent constructions in L1 (congruence between L1 and L2). Some striking results are presented in the section "Extralinguistic factors", where it is found that "the more years learners have been exposed to English in the classroom, the fewer collocations they produce in relative terms[:] increased proficiency does therefore apparently not lead to an increase in collocation use." (p. 235).

Chapter 6, "Implications of the findings", summarises the findings and presents implications both for L2 storage and processing and for teaching. In order to select suitable collocations for teaching, a three-dimensional model is introduced that relates the factors of frequency, difficulty and degree of disruption to each other. In a discussion of how collocations should be taught, a plea is finally made for the use of corpus material in the form of concordance lists.

Collocations in a Learner Corpus is a thorough piece of work. Nesselhauf arrives at results that may primarily be of interest to German teachers of English but which are also relevant for language studies in general. It is natural that her work should give rise to some thoughts and speculations.

Collocations are an intriguing part of the lexicon. If they are distinguished, as is done in this book, from free combinations and idioms, it becomes evident how difficult to handle they may be. In free combinations no restrictions apply, and in idioms there is little or no variability, but in collocations a variety of restrictions become relevant. This quality, in addition to their ubiquity, makes collocations a problem area in learner language. It might be thought that the identification and description of (correctly used) collocations in native speakers' English would be of more interest than focusing on deviant uses from a linguistic point of view, but in a pedagogical context, where one crucial aim is to find ways of teaching the correct use of collocations, the latter procedure is both natural and, as it turns out, fruitful.

The tripartite division between idioms, collocations and free combinations seems in principle to be a clear-cut one. In actual practice, however, things are different. The line between idioms and collocations is fuzzy ("some combinations bordering on the idiomatic were also considered as collocations" – p. 55)), and the delimitation of collocations from free combinations is even more problematic (*ibid*.). Nesselhauf quotes Howarth as reporting that "the actual delimitation of free combinations and collocations in verb-noun combinations proved to be extremely difficult" (p. 282). If the delimitation, even to a native speaker, is "extremely difficult", and hence not intuitively relevant, one may wonder to what extent it reflects a linguistic reality, and therefore whether it is worth making it at all. An alternative way of dealing with the problem, it seems to me, would be to regard the area of more or less fixed prefabs as a continuum with free combinations at one extreme, fixed idioms at the other extreme and collocations situated in the middle.

The author chooses to regard the elements involved in collocations as lexemes, "i.e. it is assumed that combinations such as *pay attention, pays attention, paid attention* and *attention was paid* are instantiations of the same collocation" (p. 25). Further, if an inappropriate determiner is used, such as *these* for *those* or *every* for *each*, this is disregarded (pp. 104–105). Such a policy is understandable in the presence of extensive and sometimes recalcitrant material, but it has to be pointed out that it has certain dangers. For instance, *show your hand* 'show how much power you have and how you intend to act' (Cobuild) will then be classed with the literal-meaning *show your hands*, and *stretch one's legs* 'walk about' will be classed with the literal-meaning *stretch one's leg*.

It is often mentioned in the discussion that certain collocation types are especially problematic for German students. This is a reminder to the reader that the results of the study are partly language-specific. It is also obvious that the influence of L1 on L2 is likely to be greater the more closely related the two languages are. If they are as closely related as German and English, the influence of L1 on L2 is likely to be largely beneficial, so that German learners have a head start on, say, Chinese or Arabic speakers.

At one point Nesselhauf discusses a few inappropriate collocations and hypotheses that "it seems that the learner did not use the appropriate collocation because he or she believed it to be inappropriate – in these cases probably because the collocations were considered too similar to the ones in German." (p. 226). This is doubtless an important mechanism in language learning, a mechanism that could explain the occurrence of a great many deviant forms. Traditional teaching in the contrastive tradition warns students to be on their guard against so-called false friends, so less frequent L2 expressions that are congru-

ent with corresponding L1 ones are automatically distrusted and avoided and all too often replaced with a less appropriate alternative. A recurring example in the book is the inappropriate use of *into* for *in* where German has *in*.

The judges who were asked to pronounce on doubtful examples were given the choice of clearly "unacceptable", "largely unacceptable", "questionable", "largely acceptable" and "clearly acceptable". It is not explained what criteria they were asked to use in the process. In the cases where they did not agree an average judgement was computed. The judges were also asked to supply a correct or more acceptable alternative. The task cannot have been an enviable one, as the grading of errors is most of the time a subjective affair. This will explain why the reader is not always in agreement with the judgements as presented in the book. Why, for example, is *get on your bike* appropriate and *mount your bike* inappropriate (p. 94)? And why *take a look at* for *have a look at* (p. 94), *reach a conclusion* for *come to a conclusion* (p. 95), *spend time with sb.* for *give sb. more time* (p. 153), *make fools of sb.* for *make fun of sb.* (p. 167), *give instructions on how to* for *give instructions how to* (p. 186) (both in the CobuildDirect Corpus)? In the discussion the collocations are dealt with in terms of appropriate vs. inappropriate, but the picture that emerges is rather one of constant variability. Again a representation of a cline, this time from right to wrong, comes to mind as a more adequate description of the gravity of the errors committed.

Nesselhauf touches very briefly (p. 256) on the effect of collocational deviance on communication (she uses the term "disruption"). The matter would have been worth a longer discussion. It is quite clear from the reaction of the judges that a deviant collocation need not occasion a breakdown in communication. Most of the time they understand, or think they understand, what the students are after and correct their errors accordingly.[1] Only occasionally is there a complete breakdown. But even when there is no complete breakdown, communication is nevertheless likely to be affected by a deviant language form. The hearer's attention may momentarily be diverted from the subject at hand to focus on the unexpected form. It would have been interesting, not least from a pedagogical point of view, to see just how gravely communication is disrupted by the different types of errors defined in the book.

A small point concerning the term "learner" could be worth mentioning. A learner could be anything from a beginner to a very advanced learner, and what is true of the beginner need not be true of the advanced learner. Some of the discussion, e.g. that on p. 185 about learners' use of collocations, is therefore beside the point.

The book has been carefully done, with few lapses and misprints.[2] It contains a wealth of material and many insightful and stimulating discussions.

Whether the author's recommendations for the teaching of collocations to L2 learners will be followed by the teaching community is not, of course, her responsibility. But if they are, the results will be extremely interesting to watch.

Notes

1. In one footnote it is said that "it cannot be entirely excluded that what was produced was actually intended"! (p. 292)
2. If, however, there should be another edition, it might be useful to have them seen to; a list is therefore given here:

•Slips

x, l. 15; p. 75 ll. 4, 7, 10 - occurences	*for* occurrences
46, l. 18 - only in Chapter 5 .. statistical tests will	*for* ... will statistical tests
50, l. 21 - a orthographical	*for* an orthographical
139 - underlie no restrictions	??
158 mid - human	*for human*
180, l. 1 - in neither study this influence is quantified	*for* in neither study is this influence quantified
189 mid - discernable	*for* discernible
200, l. 9 up - 63.3	*for* 63.6
212, l.2 up - probably reason	*for* probably the reason
227 mid - den *Spaß*	*for den Spaß*
233, l.1 up - seem	*for* seems
257 (figure) - distruption	*for* disruption
286 N39 - do	*for* does
287 N50 - use	*for* uses
298, l. 1 up - '+'	*for* '?' (cf. p. 52)
304 N14 - fifth	*for* fifths
313, l. 6 Kommunkative	*for* Kommunikative

•Symbols and abbreviations

x, ll. 8, 11 - if L34 means 'learner number 34', does not L2 mean 'learner number 2'?

x - A, C, O, P, V not explained

References

Cobuild = Sinclair, John *et al.* (eds.). 2003. *Collins COBUILD Advanced Learner's English Dictionary* 4th ed. London: HarperCollins.

CobuildDirect, cf. Sinclair (1987).

Sinclair, J. M. (ed.). 1987. *Looking up. An Account of the COBUILD Project in Lexical Computing.* London and Glasgow: Collins.

Ute Römer. *Progressives, patterns, pedagogy: A corpus-driven approach to English progressive forms, functions, contexts and didactics* (Studies in Corpus Linguistics 18). Amsterdam and Philadelphia: John Benjamins, 2005. ISBN 90 272 2289 4. Reviewed by **Erik Smitterberg**, Stockholm University.

Despite the wealth of existing research on the progressive in English, there appears to be no real consensus on what the central uses and functions of this feature are. Römer links this state of affairs to the fact that many previous studies lack "a broad empirical basis" (p. 1). One of the aims of her study is therefore to provide "a detailed synchronic empirical account of progressive verb forms in contemporary spoken British English" (p. 1). Römer also aims at describing the distribution of the progressive in German EFL coursebooks. A third aim is to compare these two genres with respect to the use of the progressive. Finally, based on the results of this comparison, she aims at developing "a new concept of teaching the English progressive – a concept which takes empirical findings into account" (p. 3).

Progressives, patterns, pedagogy is divided into eight chapters, which I will summarize in the order in which they appear in the book. I will then provide a critical evaluation of the study. In addition to presenting the aims of the book, Chapter 1 clarifies the scope of *Progressives, patterns, pedagogy* as regards the progressive structures included in the counts. The study covers present, past, present perfect, and past perfect progressives; other tense forms were excluded owing to their low frequency.[1] Römer also introduces the corpus-driven method of analysis she has applied throughout the study and follows this with an explanation of the structure of the book. Her corpus-driven approach centres on a close and open-minded examination of primary data that precedes the formula-

tion of categories and models (with the exception of some traditional categories such as "verb").

In Chapter 2, Römer discusses the theoretical frameworks that have informed her study. The first section is devoted to corpus-driven linguistics (henceforth "CDL"). Römer argues that CDL can be "more than a methodology, a domain of study, a theory even" (p. 7) because new theories may be "the logical result of using a new method on a new type of data" (p. 8). She also discusses the differences between CDL and corpus-based linguistics, based on how she defines these concepts. These differences include the facts that the corpus is given a more central part in CDL, and that CDL practitioners are more likely to change their theories if the theories do not fit the data. In the second section, Römer addresses "[c]ontextual approaches to the study of language" (p. 11), with special reference to the work of John R. Firth and John McH. Sinclair. The third and final section treats pedagogic and didactic grammars. Römer aims at working "towards a corpus-driven communicative didactic lexical grammar of English progressives" (p. 17), which focuses on corpus findings, spoken language, successful communication, and actual language use.

Chapter 3 chiefly comprises a selective overview of previous research on the progressive. In section 3.1, Römer discusses terms such as 'the progressive' and 'aspect', as well as accounts of the basic meaning of the progressive. In section 3.2, she argues that the increase in the frequency of the progressive in 20th-century English makes it important to describe its use accurately. The third section is devoted to how the progressive is described in two theoretical studies – Comrie (1976) and Williams (2002) – while section 3.4 deals with the treatment of the progressive in four grammars: Quirk *et al.* (1985), Biber *et al.* (1999), Mindt (2000), and Huddleston and Pullum (2002). In the fifth and last section, Römer accounts for "[p]revious empirical findings on the use of the progressive" (p. 31), with special reference to the frequency and functions of the construction, and to the contexts in which it occurs. She concludes that the works considered "differ remarkably with respect to what they actually analyse when they examine collocates of progressive forms and with respect to the type of data they use as their analytic basis", and that the results reached are thus "very different and largely incomparable" (p. 36).

Chapter 4 focuses on the progressive in spoken British English. Römer states her reasons for choosing the British National Corpus spoken subcomponent (henceforth "BNC_spoken") and the spoken British part of the Bank of English (henceforth "BoE_brspok") as sources of data, one of these reasons being that spoken British English is the main target variety for German learners. In section 4.2, Römer discusses her methods of data identification and analysis.

Based on the high frequency of their infinitive forms in BNC_spoken, 100 verbs were selected for inclusion in the study. A maximum of 200 *-ing* forms of each of these verbs in BNC_spoken and a maximum of 100 in BoE_brspok were retrieved, with a concordance window of 200 characters for each *-ing* form. Retrieved *-ing* forms that did not form part of progressives were then excluded from the counts, as were doubtful cases.[2] The resulting 9,468 progressives were imported into a database in MS Access, where they were classified on a number of parameters.

The results for spoken British English are given in sections 4.3–4.5; owing to limitations of space, my discussion of Römer's results will be selective in this review. Section 4.3 is devoted to contextual features of progressives: tense forms, contracted vs. non-contracted forms of auxiliaries, typical subjects and objects, co-occurrence with prepositions, negation, *if*-clauses, and relative clauses, occurrence in questions, and adverbial specification. The two corpora, for which separate figures are presented in sections 4.3 and 4.4, usually yield very similar results: for instance, progressives in both corpora favour contracted forms of auxiliaries and personal pronouns as subjects. In section 4.4, Römer considers the functions of the progressives in her material. Among other things, she identifies two common function-related features of the progressive, continuousness and repeatedness, of which the latter has not received extensive treatment in previous research. The most frequent combinations of these features, 'continuous + non-repeated' and 'continuous + repeated', are singled out as the two central functions of the progressive. She also identifies seven additional functions, which may co-occur with either of the central functions: 'general validity', 'politeness or softening', 'emphasis or attitude', 'gradual change and development', 'old and new habits', 'framing', and 'shock or disbelief'. Unlike the two central functions, many of these additional functions occur chiefly with a restricted set of verbs. Section 4.5 covers the same contextual features and functions as were treated in sections 4.3–4.4, but now the results are presented as per main verb.[3] Römer's results show that progressives of different main verbs frequently pattern quite differently with respect to the features and functions investigated: for instance, progressives of some main verbs occur chiefly in non-continuous situations despite the centrality of continuous situations to the functions of the progressive. She concludes that these results imply "an apparent need to question the existence of a purely grammatical progressive in favour of a lexical-grammatical one" (p. 169).

Chapter 5 focuses on the progressive in EFL coursebooks. The first section of the chapter is devoted to the question of why the progressive constitutes a problem for German learners; Römer argues that the lack of a similar, equally

grammaticalized construction in the learners' first language and "inadequate descriptions of language phenomena in teaching materials" (p. 173) may be important factors in this regard. The following three sections discuss Römer's compilation of a German EFL coursebook corpus and the retrieval and classification of data from this corpus. Two recent and widely used series of coursebooks for learners between the ages of ten and sixteen were selected: *Learning English Green Line New* (henceforth "GLN") and *English G 2000* (henceforth "EG 2000"). Passages that represented speech in the books were scanned and converted to text files. The data were retrieved and classified in a manner similar to that used for spoken British English, but in this case all progressives of the 100 selected main verbs were included, resulting in a total of 702 progressive verb phrases.

Römer's results are given in sections 5.5–5.7: section 5.5 focuses on contextual features, 5.6 on functions, and 5.7 on lexical patterns in the distribution of contextual features and functions. As regards many of the contextual features, there are greater differences between the two coursebook series than were attested between BNC_spoken and BoE_brspok. For instance, contracted auxiliaries in progressive verb phrases are much more frequent in EG 2000 than in GLN. The distribution of functions of the progressive is more similar in the two coursebook series; most additional functions, for example, are rare in both series. Given the differences attested between the two coursebook series, Römer treats them separately in the lexical analysis in section 5.7. A great deal of variation according to main verb is attested as regards the distribution of progressives across contextual features such as tense form and subject. In addition, the distribution of progressives of the same main verb with regard to a contextual feature often differs substantially between the two coursebook series; for example, only progressives of DO are frequently specified by adverbials of place in GLN, while in EG 2000 such frequent specification occurs with more verbs. In the last section of Chapter 5, Römer addresses the presentation of the progressive in the textbooks, grammars etc. that form part of GLN and EG 2000. The two series are quite similar with respect to the order in which different types of progressives are introduced (present – past – present perfect – past perfect), an order which Römer considers reasonable against the background of the results presented in Chapter 4. However, she argues that "a rather simplified picture that deviates quite a bit from actual usage" is presented as regards the functions of the progressive, and calls for "the inclusion of an enhanced lexical-grammatical perspective" on the progressive (p. 241), and for more information on collocational patterns.

Chapter 6 is devoted to a comparison of the results reached in Chapters 4 and 5, with the aim of assessing the extent to which the use of the progressive in the coursebooks mirrors spoken usage by native speakers. Römer considers contextual and functional features in separate sections; comments on lexical-grammatical relationships are provided in both sections. A great many differences between either or both of the coursebook series and spoken British English are attested; for instance, GLN has a lower proportion of contracted auxiliaries than EG 2000, BNC_spoken, and BoE_brspok. There are also differences on the functional side, such as progressives in non-continuous and repeated situations being rare in both coursebook series (as are progressives with most of the additional functions identified). A succinct summary of the ways in which the coursebooks deviate from spoken British English usage concludes the chapter.

In Chapter 7, Römer addresses the pedagogical implications of her results. She argues for moderate changes to the coursebooks so that their texts become more representative of native-speaker usage, and claims that CDL has the potential to improve materials design in this regard. She also argues that coursebooks should be based on authentic, unedited language rather than invented texts or authentic but edited language, and that learners' communicative competence would be increased if coursebooks first focused on central aspects of linguistic features, postponing the treatment of more marginal aspects. A closer integration of grammar and lexis is also called for, so that progressives of a given verb are presented with the functions and in the contexts with which they are typically associated in native-speaker usage. The desired outcome of making these changes would be the type of corpus-driven, communicative, didactic, and lexical grammar that Römer introduced in Chapter 2 and returns to here. She also discusses where in the learning process learners should be presented with each relevant contextual feature, function, and tense form.

The eighth and last chapter of *Progressives, patterns, pedagogy* summarizes Römer's findings. It also addresses some limitations of the study and outlines some areas where more research, as well as more work in the field of corpus compilation, is needed.

Progressives, patterns, pedagogy is truly impressive in scope. The analysis covers more than 10,000 progressives, all of which have been classified on a large number of parameters. The wealth of data means that many of Römer's results are reliable from a quantitative perspective. However, Römer also succeeds in making several interesting and valuable observations regarding smaller groups and single examples of progressives throughout Chapters 4–6. As most existing research is based predominantly on written texts, her decision to analyse spoken English further adds to the novelty of her approach. Of particular

value is her decision to use two corpora of spoken English and compare their distribution of progressives in order to achieve the "highest possible generalisability of the findings" (p. 43). The inclusion of coursebook material is another original feature of the book, as are the explicit comparison of the two genres covered and the implications for materials design raised by this comparison. The lexical perspective adopted is also praiseworthy; it adds significantly to our knowledge of the progressive and enables Römer to make a strong case for the inclusion of more lexical information in teaching materials. *Progressives, patterns, pedagogy* contains a great many illustrative graphs that help the reader to interpret the results, and Römer also provides a wealth of corpus examples, a sizeable References section, and an Index. As regards Römer's text, the sections on material selection and corpus compilation stand out as particularly lucid. Overall, my impression of *Progressives, patterns, pedagogy* is thus highly positive.

Nevertheless, there are a few areas where improvement seems possible. First, in a few cases, I would have appreciated a more detailed account of the criteria Römer used to identify and classify her data. One of these cases concerns adverbial specification. If I understand Römer's account correctly, the *if* in the example *Then, after looking around to see if anyone was listening, she went on* (p. 222) is classified as an adverbial specifying and co-occurring with the progressive *was listening*. However, in many traditional accounts, *if* would be classified as a subordinator in this sentence, and subordinators are not normally regarded as one of the structures that can constitute adverbials (see e.g. Quirk *et al.* 1985: 489). Römer may be using "*if*" as a short form for '*if*-clause', but as the clause *if anyone was listening* would typically be classified as nominal rather than adverbial in this example, the classificational framework applied would still be unclear to me.[4] (In addition, even in a sentence such as *We may be in trouble if anyone was listening*, where *if anyone was listening* would be an adverbial clause in traditional accounts, it is arguable that the clause co-occurs with the progressive *was listening* but does not specify it, since the progressive forms part of the adverbial clause.) Römer's classification may of course differ from that found in traditional accounts and may well be perfectly defensible, but she would have increased the replicability of her results by overtly stating the criteria she applied when classifying her data on every parameter included.[5] This holds regardless of whether these criteria originated in pre-existing frameworks or emerged from close examination of the data. Such a statement might also have helped to explain some differences between Römer's results and those of previous research, since the differences may be due in part to different classificational frameworks being applied to the data.

Secondly, after presenting results for spoken British English and coursebook texts in Chapters 4 and 5, Römer has to repeat many of these results in the shape of combined tables and figures in Chapter 6, where the two genres are compared. It might have been possible to conflate Chapters 5 and 6 in order to remove some of this repetition, as Römer already comments on several differences between the coursebooks and spoken English in Chapter 5.

Thirdly, I would have appreciated more information regarding the application of the chi-square test for statistical significance. Römer states that the test was applied, and the results commented on, "whenever considered appropriate" (p. 60), but it is not always clear to me why a given difference between samples was not tested for significance. More detailed information on degrees of freedom, chi-square and p values, etc. as regards the tests that were carried out would also have been welcome.

Finally, although *Progressives, patterns, pedagogy* comes across as well-edited overall, there is one apparent usage problem in the book. An *-s* occurs after the apostrophe in a great many genitive plurals, e.g. "researchers's" (p. 8), "scholars's" (p. 113), "editors's" (p. 175), and "learners's" (p. 296) instead of the expected forms *researchers', scholars', editors',* and *learners'*.[6]

In sum, then, *Progressives, patterns, pedagogy* is an original and praiseworthy addition to existing research on the progressive and on how the construction is presented to learners. It is hoped that Römer's successful endeavour will inspire further research in these areas, which will both add to our knowledge of spoken English and help to create more suitable teaching materials.

Notes

1. In addition, as Römer makes clear in Chapter 4, limitations were imposed with regard to the number of main verbs included in the study and the maximum number of progressives included per main verb.
2. However, on p. 115, Römer discusses a potential progressive of the verb SUPPOSE that appears to have been included in the counts even though "it is not clear whether *supposing* is really part of a progressive construction" in the example.
3. BNC_spoken and BoE_brspok are treated together in section 4.5, as "they showed largely similar distributions in almost all context and function categories" (p. 112).
4. This may be what Römer means when she discusses "non-conditional *if* (in the 'whether' sense)" (p. 79). However, she nevertheless classifies this item as belonging to the group "other adverbials".

5. It is of course possible that Römer does discuss both the criteria she used to define adverbial specification and the reference of *if* (and of similar items that she subsumes under the label "adverbial"), and that I have failed to notice this. Römer discusses many of her criteria admirably, but often does so in different places in the book, which may make it more difficult for readers to access the information; for instance, she elaborates on what she subsumes under the label "preposition" in note 59 (p. 303), and comments on her inclusion of *be going to* future expressions as progressives of GO in a discussion of time reference (pp. 154–155). (In contrast, the additional functions of the progressive are described and defined the first time Römer presents quantitative information on their distribution in section 4.4.5, which makes perfect sense.) Giving such discussions more prominence in the running text as well as in the Index and/or Table of Contents would clearly have added to the value of *Progressives, patterns, pedagogy.*
6. Although forms such as *researchers's* do not occur in descriptions of Standard English usage in works such as Quirk *et al.* (1985) and Biber *et al.* (1999), they are occasionally found in English texts. A Google search restricted to the .edu domain for the form *researchers's* returned 64 matches (the search was carried out on 13 January, 2006, and included English-language material only). However, in some of these cases the context strongly suggests a genitive singular reading.

References

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. 1999. *Longman grammar of spoken and written English*. Harlow: Pearson.

Comrie, Bernard. 1976. *Aspect: An introduction to the study of verbal aspect and related problems*. Cambridge: Cambridge University Press.

Huddleston, Rodney and Geoffrey K. Pullum. 2002. *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.

Mindt, Dieter. 2000. *An empirical grammar of the English verb system*. Berlin: Cornelsen.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London and New York: Longman.

Williams, Christopher. 2002. *Non-progressive and progressive aspect in English*. Fasano: Schena Editore.

John Sinclair (ed.). *How to use corpora in language teaching*. Amsterdam: John Benjamins 2004, 299 pp. ISBN: 90-272-2283-5. Reviewed by **David Oakey**, University of Birmingham.

Few readers of this journal will need to be reminded of the COBUILD project, John Sinclair's great lexicographical contribution to English language teaching in the 1980s, which collected one of the first sizeable electronic language corpora – the Bank of English – in order to produce a new range of dictionaries for the international user of English. The Bank of English yielded further reference materials during the ensuing decade: a grammar, books of common lexico-grammatical patterns, a range of vocabulary and stylistic guides, and a series of textbooks constructed around a lexical syllabus derived from frequency data from the corpus.

Having been responsible for something approaching a paradigm shift – after which few language reference materials have been produced without prior construction of a sizeable corpus – Professor Sinclair has continued, in a variety of research publications (e.g. Sinclair 2003; 2004), to show how the examination of a well-planned and constructed corpus can bring to light new facts about language. The book under review here introduces the work of a recently arrived generation of mostly European corpus researchers and teachers, and fits in well with previous volumes in Benjamins' *Studies in Corpus Linguistics* series, e.g. Partington (1998) and Ghadessey *et al*. (2001), which have presented work at the interface between corpus research and language teaching.

The appearance of this collection is timely, since, as Sinclair points out on page 272, corpus evidence of language use is often viewed as a threat by various participants in the language education process: enormous numbers of unmediated examples can overwhelm unwary students; examples containing abstruse, highly context-dependent uses can undermine the confidence of language teachers; research findings can reveal to publishers that their materials contain misleading statements about the language.

Happily, the tone of the papers in this volume is more optimistic, and the authors take a "can-do" approach to the issues involved in both using corpora directly in language teaching classrooms, and in identifying linguistic features from corpora that are likely to be of use to teachers. The book consists of twelve chapters, each of which, while covering various aspects of the volume's overall theme, could stand alone as a self-contained paper; their introductions often cover similar ground in justifying the importance of corpora for language teaching, and their references are presented at the end of each chapter rather than

being collected at the end of the entire volume. The book is divided into four sections: two chapters on the corpus and the teacher, four on corpus resources, two on research, and four more on computing resources, although some aspects of papers in one section can often be seen to connect to the theme of another section.

The chapter by Silvia Bernadini opens the section on the corpus and the teacher, and its focus on the use of corpus data for schema restructuring is very much in the tradition of Tim Johns, whose pioneering work in this area (e.g. Johns 1991) is cited by several of the papers in this collection. The approach, named "discovery learning" by Bernadini, seems very appropriate for the cohort of student translators for whom it is intended, and "it encourages learners to follow their own interests whilst providing them with opportunities to develop their capacities and competences so that their searches become better focused, their interpretation of results more precise, their understanding of corpus use and their language sharper" (p. 23).

The next paper by Amy Tsui illustrates the use of corpus data to raise teachers' language awareness; it describes a project in which teachers' questions about the English language were answered by someone with access to a corpus. While interesting examples are given, such as on the difference between *tall* and *high*, it could be said, however, that one generalisation from the evidence in the paper is rather misleading. A teacher submitted a question regarding the use of articles:

> Should I say "When the teacher was teaching, they listened to the walkman"
> Or "When the teacher was teaching, they listened to walkman" (p. 54)

The answer given is that *walkman* is an "ordinary countable noun, and therefore the indefinite article 'a' or a possessive pronoun would be used," (*ibid.*) with 28 instances of *walkman* in the corpus apparently confirming this introspection. Tsui does not make clear what the final advice to the teacher was in the light of these results, but the generalisation seems to lead to the following possible sentences:

> "When the teacher was teaching, they listened to a walkman."
> "When the teacher was teaching, they listened to the walkman."
> "When the teacher was teaching, they listened to their walkman."

While the word before *walkman* in the above three sentences may be a grammatically possible choice in each case, the sentences also conjure up a distinctly odd image in the reader's mind. Odd-sounding invented sentences are nothing new in language teaching (e.g. Cook 2001) but they are usually constructed with the intention to highlight features of language use. The problem seems to be that juxtaposing the original pair of sentences did not highlight the article problem on which the teacher was trying to focus. I would imagine a student asking how – if two activities in a sentence are going on simultaneously – can one activity be described in the past continuous and the other described in the simple past.

The above example suggests that reference to real corpus data may fail to make a confusing invented sentence any less confusing. It probably would not have been much more useful here to look for corpus evidence for the plural of *walkman*, as this is problematic whether or not one supports a social or mentalist view of language (e.g. Pinker 1994: 143); the table below illustrates the contradictory nature of the evidence.

Table 1: Frequencies of *walkman*, *walkmen* and *walkmans*

	British National Corpus	Bank of English	Google hits
walkman	179	489	7,820,000
walkmen	3	18	1,620,000
walkmans	18	70	1,020,000

The paper by Susan Conrad, which opens the next section on corpus resources, shares the aim of the previous chapter of a focus on raising teacher awareness. Her chapter stresses the importance for language teachers of a knowledge of language variation, and points out that "we are misrepresenting language in materials that we use with students" (pp. 68–69). Her chapter contains two useful examples of corpus insights into register variation, first comparing the use of *though* as a linking adverbial in different registers with how it is taught in "general" English textbooks, and then presenting a multi-dimensional comparison of the spoken language of class sessions with a lecture in an EAP textbook.

Mauranen's paper makes a similar proposition to Conrad's that "what is taught as functional language use is not necessarily in agreement with what is frequent in the language" (p. 90). Her chapter surveys three issues: authenticity, communicative utility, and – in a lengthy third section on formulaic expressions – outlines the tension between the ease with which corpus evidence for formulaic expressions can be found and the difficulty with which they can be made to

become part of a learner's spoken language repertoire. She concludes briefly with a call for the construction of corpora which more accurately reflect the use of English around the world: "international learners are not primarily in need of British models, but a sensible range of more international varieties, including non-native expert use" (p. 104).

A shorter chapter by Pereira exemplifies the use of corpus data in revealing new evidence relating to the teaching of Portuguese. Particularly striking is the data on the frequencies of different inflections across disciplines and between spoken and written modes.

Nesselhauf devotes her paper to an in-depth survey of learner corpora, their potential and limitations, and concludes with an extended example of the use of data-driven learning with learner corpora. The difference between this and the "discovery learning" in Bernadini's chapter is that with learner corpora the students are learning from negative evidence. As Nesselhauf points out on page 141, this has intriguing implications for learner motivation. Indeed, I have noticed that my own students are more than eager to spot other learners' mistakes, and that they find doing so infinitely preferable to identifying their own!

Gyula Tankó's chapter provides a longer example of the advantages to be had from comparing a corpus of learner writing with one of native writing, and presents a detailed piece of research which amounts to a case study on the use of adverbial connectors in argumentative writing. Studies such as this take native use of a language as the norm in relation to which learners "overuse" and "underuse" various features. When contrasted to Mauranen's call for more "lingua franca" corpora containing instances of language use by non-native users, Tankó's approach illustrates the continuing difference of opinion as to what constitutes an appropriate linguistic target.

The chapter by Ute Römer concludes the section on research with another comparative study, this time between how modal auxiliaries are taught in German EFL textbooks and how they are used in British English corpora. Her study found discrepancies between corpus and textbook data, and it is a strength of her work that her recommendations are not just about *what* language should be taught but *when:* she suggests changing the order in which modals are introduced in the syllabus so that more frequent verbs are taught before less frequent ones.

The next two papers concentrate on the technical aspects of retrieving and manipulating electronic corpus data. Michael Barlow's chapter takes the reader through the various perspectives from which a corpus can be viewed, and points out the different features which can be noticed as software tools present the language data in more and more abstracted representations, i.e. from concordance

lines to wordlists, and on to collocate tables and lexical frameworks such as *the _____ that*.

Pernilla Danielsson's paper is rather different from the others in this collection in that it provides actual programs, written in the "Perl" programming language, with which simple corpus-tidying operations such as tokenising and splitting can be performed. Like many computer users who are not programmers, I long ago adopted a mouse-and-icon graphical interface with my computer, and I have not used a command line since the early 1980s. It is interesting to take back control; at times using these programs I felt like a character from Isaac Asimov's story who – in a futuristic world where computers have for long performed all actions for humans – suddenly rediscovers mental arithmetic, and is amazed by the feeling of power he has. It was also chastening to see once again the catastrophic effect on the operation of a program of a single missed space or curly bracket. While some of the explanations in Danielsson's chapter are hard to follow, e.g. "there is no need not to save the file between each individual command here. Instead use the pipe | to combine them together," (p. 242), her breezy, can-do approach, e.g. "all that is needed to rectify each of these difficulties is a small program," (p. 227), carries the reader along.

Pascual Pérez-Paredes, in a wide ranging chapter, places the use of corpora in language teaching within the wider Computer Assisted Language Learning (CALL) context, and looks ahead to the possibilities offered by making learner oral corpora accessible on networked computers in language laboratories.

In the final chapter John Sinclair examines four problematic areas of language and, in extended case studies, shows that each are pseudo-problems: ambiguity – a result of inappropriate theorising of language; variation – not as complicated as first thought; terminology, which can be made less misleading; and incompleteness of description, which can be controlled through focusing on frequent occurrences.

In all this volume contains something for everyone; there is much in its 300 pages to interest teachers and researchers alike.

References

Asimov, Isaac. 1958. The feeling of power. *If: Worlds of Science Fiction*, 58(2): 4–11.

Cook, Guy. 2001. 'The philosopher pulled the lower jaw of the hen.' Ludicrous invented sentences in language teaching. *Applied Linguistics* 22(3): 366–387.

Ghadessy, Mohsen, Alex Henry and Robert L. Roseberry (eds.). 2001. *Small corpus studies and ELT: Theory and practice*. Amsterdam: John Benjamins.

Johns, Tim. F. 1991. From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. *ELR Journal* 4: 27–45.

Partington, Alan. 1998. *Patterns and meanings.* Amsterdam: John Benjamins.

Pinker, Steven. 1994. *The language instinct*. London: Penguin.

Sinclair, John. McH. 2003. *Reading concordances: An introduction*. London: Pearson Longman.

Sinclair, John. McH. 2004. *Trust the text. Language, corpus and discourse*. London: Routledge.

Erik Smitterberg. *The progressive in 19th-century English. A process of integration*. Amsterdam and New York, NY: Rodopi, 2005. xvi + 284 pp. ISBN 90-420-0515-7. Reviewed by **Ute Römer**, University of Hanover.

Whoever decides to carry out research on the English progressive will note that there is a wealth of literature available that deals with this form, its development, distribution and functions. What has been missing so far, however, is a detailed diachronic account of and a systematic cross-genre approach to the use of progressives in the 19th century. Smitterberg's monograph addresses both research desiderata in that it investigates the development of the construction between 1800 and 1900 and examines its distribution across genres.

The book is a revised and shortened version of the author's doctoral thesis, presented at the University of Uppsala in 2002, which discusses the results of a detailed analysis of 2,440 progressive form tokens taken from *A Corpus of Nineteenth-Century English* (CONCE). The volume consists of eight chapters which are preceded by a detailed table of contents, a list of tables and figures, and a preface, and followed by a comprehensive reference list and three appendices with information about the corpus material and on statistical tests carried out on the data. Unlike other volumes in Rodopi's Language and Computers series, the book does not have an index.

Chapter 1 ("Introduction") describes the background of the study, delimits its scope, and clearly sets out its aims. In addition to his general and primary

aim, i.e. to account for the use and development of the progressive in Late Modern English, Smitterberg defines four secondary objectives: (1) to address methodological issues in corpus-based linguistics, focussing on different frequency measures of the progressive and statistical considerations, (2) to analyse the relations between progressive use on the one hand and time, genre, and writer gender on the other, (3) to examine the type and structure of progressive verb phrases, and (4) to investigate the co-occurrence of the progressives with other linguistic features, such as temporal adverbials. The chapter also discusses some central terminological issues and describes the analytical frameworks of the study, as there are what the author labels "the general framework of corpus linguistics" (p. 13), the variationist approach, and multi-dimensional analysis in Biber's (1988) sense.

Chapter 2 ("Material and data") gives a clear and comprehensive presentation of the CONCE corpus which serves as the empirical basis of Smitterberg's study. CONCE is a corpus of 19th-century English texts of roughly one million words. It provides follow-up material to the well-known Helsinki Corpus and was compiled with a cross-genre perspective in mind. The chapter also describes the procedures of data retrieval from the corpus. This data retrieval was followed by a time-consuming manual filtering and intensive post-processing in order to separate progressives from competing forms, such as adjectival participles or gerunds. These steps seem to have been carried out extremely carefully so that the result is a valuable set of data. One problem I see with respect to achieving the highest possible level of representativeness, however, is related to the sampling policy adopted in the compilation of CONCE. Since linguistic features are not distributed evenly across texts, any form of sampling bears a risk.

The focus of Chapter 3 ("The frequency of the progressive in 19th-century English") is mainly methodological. Smitterberg devotes a large part of the chapter to the question how best to measure the frequency of the progressive and discusses four possible ways, two of which (the M- and S-coefficients) are then applied to the CONCE data. The application of the M- and S-coefficients shows how drastically different methods of calculation can affect the results of a corpus study and the interpretation of the findings, and hence calls for highest caution in any type of quantitative language research. The author goes on to analyse how far the use of the progressive in the 19th century was affected by the time of writing, the text genre, and author or addressee gender. He also defines the concept of "integration", a label Smitterberg prefers over the – in his opinion – narrower term "grammaticalization".

Chapter 4 ("M-coefficients and factor score analysis") discusses the application of a multi-dimensional analysis to the formal and functional distribution of the progressive. To achieve this, Smitterberg compares the M-coefficients he determined in Chapter 3 (capturing the distributional dependencies of progressives on time, genre and gender) with dimension scores that were the results of a factor-score analysis carried out by Christer Geisler (see e.g. Geisler 2002) – an analysis which has highlighted some important co-occurrence features of the progressive. The correlations Smitterberg hence identifies (for example that between high frequencies of progressives and features that are typical of involved production) are shown to be highly genre-dependent. From these findings the author goes on to deduce that the progressive and the features on a correlating dimension express similar functions. He thus links frequency and functional accounts of the construction and interprets them in the light of the concept of grammatical integration.

Further indicators for an increasing degree of integration of the progressive in English grammar are dealt with in some detail in Chapters 5 and 6. Chapter 5 ("Morphosyntactic variation in the verb phrase") takes a closer look at the patterns which the progressive commonly forms within the verb phrase and examines whether and how the degree of integration of the form can be determined by means of co-selections with the formal features of tense, the perfect, active/passive voice, and modal auxiliaries. Some diachronic trends are observed for selected genres, for example an increase of present tense progressives in Fiction or an overall decrease of progressives with modal auxiliaries. On the whole, complex verb phrase patterns were found to be rather rare in the CONCE data and did not become more frequent in later stages of the 19th century.

The next chapter, Chapter 6, is also dedicated to variation in the use of the progressive, this time to "Variation with linguistic parameters". In his selection of parameters that can affect the distribution of the form, Smitterberg solely relies on the findings of previous research, which may be considered problematic from the perspective of corpus-driven research, i.e. research which puts the corpus in pole position and, instead of applying existing theories or frameworks to the data, aims to work towards new systems that accept and reflect the evidence. Even though there is no guarantee that a more data-oriented, less theory-driven approach may not have highlighted other, perhaps even more revealing parameters, the variables Smitterberg selects (main verbs in the progressive, situation types, agentivity of subjects, temporal adverbial specification and clause types) definitely provide some very interesting insights into the phraseological or co-selectional behaviour of progressives. The analyses also nicely demonstrate on which of the parameters the progressive shows clear tendencies

towards diachronic change. The author thus skilfully relates his findings to the overarching issue of integration.

The results of a final analysis of the CONCE progressives are discussed in Chapter 7 ("The not-solely-aspectual progressive: An analytical approach"). While Chapters 3 to 6 were mainly focussed on the more frequent aspectual functions of the progressive, special attention is now given to three types of progressives that express something beyond pure aspect and which prove to be relevant to the integration of the form in Late Modern English: (i) progressives modified by adverbials like *always*, (ii) potentially "experiential" progressives, and (iii) interpretative progressives. Worth mentioning in this context is Smitterberg's careful analysis and meticulous functional annotation of the forms he had retrieved from CONCE, which must have been extremely time-consuming. Like the previous chapters, Chapter 7 provides an extensive survey of the literature and is interspersed with numerous references to relevant studies.

So far, a number of apparently separate analyses have been carried out on the extracted CONCE progressives. The final chapter, Chapter 8 ("Concluding discussion"), interrelates the findings of these analyses and reminds the reader of their common aim: to explore the process of integration of the progressive into the grammar of Late Modern English. A summary of the major empirical findings of the study clearly indicates that the historical development of the progressive is strongly genre-dependent and that it can be considerably misleading to postulate a general increase of the progressive in English. What also becomes clear is that integration is a highly complex process that requires extensive investigations based on larger amounts of language data from different periods and text types, and including a number of formal and functional features in the context of the construction under scrutiny. Smitterberg is certainly aware of this. He points out that there is a need for further studies that look into a wider range of co-selection phenomena of the progressive, its relation to other verbal constructions, sociolinguistic variation, and genre development. Throughout the book, the author is rather critical of his own work and refers to the limitations of the study – perhaps a little too often. *The progressive in 19th-century English* is a very detailed and thorough study, but by using so many disclaimers and references to limitations and caveats, Smitterberg runs the risk of conveying a different impression to the reader. Concerning the overall design of the study, I would have wished the author had been a little more "radical" in approaching the data instead of relying almost exclusively on previous accounts in analysing the contexts and functions of the CONCE progressives. However, the points of criticism mentioned in this review should not detract from an inspiring and

insightful book that makes valuable contributions to the fields of diachronic linguistics, syntax, and genre analysis.

References

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

Geisler, Christer. 2002. Investigating register variation in nineteenth-century English: A multi-dimensional comparison. In R. Reppen, S. Fitzmaurice and D. Biber (eds.) *Using corpora to explore linguistic variation*, 249–271. Amsterdam: John Benjamins.

Martin Wynne (ed.). *Developing linguistic corpora: A guide to good practice.* Oxford: Arts and Humanities Data Service & Oxbow Books, 2005. 87 pp. ISBN 1-842-17205-0. Reviewed by **Hans van Halteren**, University of Nijmegen.

When I was asked if I wanted to review this book, I had a quick look at the cover text. What I read was that this book would tell its readers how to properly build linguistic corpora. Now this would be a dearly needed book, since many corpus users these days are more likely to just mess around than follow thorough methodology. I was also a bit sceptical, though, since 87 pages did sound somewhat short, but of course I was hoping this was due to succinctness and clarity. What did I find when actually reading the book?

After a short introduction by the editor, the first chapter, by Sinclair, takes us through all the important issues one has to consider when designing and compiling a corpus. Projected uses and users, criteria, sampling, representativeness, balance, topic, size, homogeneity, all are addressed. And here it is also made clear that, as far as this book is concerned, the only purpose of corpora is the study of language. In Chapter 2, Leech describes what analytic annotation can be added to a corpus. The main example is tagging, but all other possible layers are present. Burnard follows with a chapter on meta-data, describing all those other things you should keep records of, and how you can put the information in a useful and standardized form. Chapter 4 concludes the annotation and storage

chapters, as McEnery and Xiao provide ample information on character encoding. In Chapter 5, Thompson informs us about spoken language corpora. We are taken through the whole pipeline of collection, transcription, annotation and distribution, and some of the important issues are addressed in more detail. The last chapter is by Wynne himself and discusses what to do once the corpus is ready, such as archiving and distribution. The book is completed with an appendix in which Sinclair gives some more hints on actually building a corpus. There is no index or glossary, and the references appear limited to what the authors mentioned in their chapters, but the book has already made clear it assumes its readers know how to use the internet and can find more information themselves.

Are all the necessary topics sufficiently addressed? Well, mostly yes. Maybe copyright issues might have deserved some more room, as might parallel/comparable corpora, annotation in tiers, the issue of lexical/potential versus contextually appropriate annotation, and/or syntactic analysis with dependency schemes. What I find especially lacking is the topic of software which can be used for the development and exploitation of corpora. As the book correctly points out, the best software is much more subject to change than the principles of corpus methodology. Still, the topic might have been addressed in a more general way, for example by describing how your choices of software might influence your work or pointing out that it could be useful if you archive/distribute your software with your corpora.

A personal point of criticism is that the book is very much focused on English – perhaps even British English. Remarks about other languages are only rarely made (albeit in most of the right places). In Chapter 6, where my impression of English-centeredness really grows too large to ignore, I also get an inkling of a possible reason. Here, the AHDS (Arts and Humanities Data Service) turns out to be not only the publisher, but also the preferred place to store your corpora. Still, even if this book may have been written with British researchers in mind, most of its contents are valid, or at least constitute a good enough starting point for foreigners too.

The unavoidable local wrinkles are of course also present in this book. Leaving aside the editor's own chapter, spelling and grammar errors are rare. Unfortunately, they are often such that they force you to reread and interpret what is really meant, as do the floating figures on page 64, but in each case I could reconstruct the meaning. As for content, I found only the following two things strange enough to mention.

On page 5, Sinclair spends half the page to argue that mark-up, if it has to be included at all, should be kept out of the actual text. This remark seems a bit out of place here, seeing that Chapters 2 and 3 will be about mark-up; moreover, it is

difficult to figure out what this argument is doing in the section on criteria for the selection of texts to include in your corpus. This is not the only place where the authors could have created a better book by reading more closely what the others are already saying in their chapters (e.g. the re-arguing that it is advisable to use TEI in Chapter 5 and the deviant section marking in Chapter 3).

The second local wrinkle (or more than that) is the fact that Thompson's references seem a bit old. All but one date from 1998 or earlier. And the one more recent one is a very unfortunate reference to Meyer (2002), where it is claimed that if you want to syntactically analyse your spoken corpus, you will have to normalize your text, which should be anathema to corpus linguists. For spoken corpora, I would refer the reader also to some more recent literature, e.g. that on the Spoken Dutch Corpus or the Santa Barbara Corpus of Spoken American English. Here you might find examples of issues like stand-off annotation, coordination of multiple annotation layers (and multiple annotating research groups) and error correction procedures, all of which might have received some more attention in this guide.

After the local wrinkles, I get to my main disappointment. Contrary to my expectations, the book does not tell me how to do all the things I need to do. For design and compilation, Sinclair lists all the issues, gives arguments, and then tells me to use my own common sense. His appendix does not help either, as it is more about speeding up things, even cutting corners and rationalizing this, than about the basic choices. For analysis, Leech lists all the issues, and then for each gives me pointers to the literature. Thompson's spoken corpora are a mix of these two, but again I do not get helpful instructions. Burnard is more explicit (use TEI), as are McEnery and Xiao (use Unicode, UTF-8) and Wynne also gives good instructions.

But then, my disappointment turns out to be based on my own preconception. When reading the cover text more closely, I see that I am not promised a "how to build a corpus in ten easy steps". I am promised an exploration of the key issues and pointers to more information. And this is what the book delivers. So if you are in need of a step-by-step manual, you will have to wait a bit longer. But if you want to know what to pay attention to, and are willing to think and search further yourself, then this book is a useful one to have on your shelf.